

©2020

William Alexander Hansen

ALL RIGHTS RESERVED

MULTI-SCALE PROTEIN DESIGN UTILZING SYMMETRY

by

WILLIAM ALEXANDER HASNEN

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Quantitative Biomedicine

Written under the direction of

Sagar D. Khare

And approved by

---

---

---

---

New Brunswick, New Jersey

May 2020

# ABSTRACT OF THE DISSERTATION

Multi-scale Protein Design Utilizing Symmetry

by WILLIAM ALEXANDER HANSEN

Dissertation Director:

Sagar D. Khare

Despite significant advances in the field of computational protein modeling and design, the prediction of *de novo* metal-coordination and supramolecular assembly remains a largely unexplored area of bottom-up design. The computational tools and methods outlined in this dissertation are intended to reduce design complexity, promote a generalizable design framework, and lay the groundwork for the development of *de novo* metal-coordination and supramolecular assembly.

Thirty percent of proteins in Nature, by estimation, contain metal binding sites and these exhibit a diverse array of structural and functional utility. Among them, multi-nuclear metal clusters perform the most exquisite chemistry such as water oxidation, hydrogenation, and nitrogen fixation. In order to harness the catalytic potential of multi-nuclear metal clusters, we propose a general method for the design of multi-nuclear metalloprotein protein precursors, one that exploits the benefits of symmetric coordination and polydentate non-canonical amino acid derivatives. We have developed a computational searching algorithm (SyPRIS) to locate within a library of structurally determined symmetric protein oligomers a constellation of backbone atoms with a geometry compatible with a desired metal cluster. SyPRIS is shown to have 100% accuracy in the prediction of the native metal-binding sites of known symmetrically coordinated metal ions at the interface of oligomeric proteins ( $C_2$  and  $C_3$ ). Furthermore,

in a crossmatch study of the benchmark structures, more than 1000 novel metal binding sites with native-like scores are predicted, suggesting the utility of SyPRIS for the incorporation of non-native amino acid coordination of a desired complex.

In order to complement the benefits that symmetry offers for reducing design complexity, we sought to expand the palette of available biocompatible non-native amino acid derivatives. A two-step synthesis provides a high-metal affinity bioconjugatable unit, 2,2'-(ethene-1,1-diyl)bis(1-methyl-1H-imidazole) or BMIE. Direct attachment of BMIE occurs by thiol-selective conjugate addition on the surface of a carboxypeptidase G2 variant (S203C). Additionally, we find that BMIE adducts can bind an assortment of divalent metal ions (Co, Ni, Cu, and Zn) in various bi- and tri-dentate tetragonal coordination geometries. Non-BMIE coordinated positions of the copper-bound modified protein display lability in the presence of several counter ligands ( $\text{H}_2\text{O}/\text{OH}$ , tris, and phenanthroline), which highlights the potential for future catalytic applications. The site-selective modification of proteins for high-affinity metal-binding, combined with the ease of adduct formation and metalation make BMIE an attractive tool to augment multi-nuclear metalloprotein design.

The design of protein-based assembly is a burgeoning field with applications in biomedicine and bioremediation. However, topologies have been limited to integer-dimensions. Additionally, many questions remain with respect to the impact of protein anisotropy, colocalization on catalytic pathways, and the effect of kinetics on self-assembly. In order to address these unexplored questions, we developed a general design method and computational tools for fusion-mediated protein assembly directed by symmetry. To show that our design method could be extended to any set of symmetric

proteins, we chose two members of the atrazine degradation pathway that are known to be symmetric: AtzA ( $D_3$ ) and AtzC ( $D_2$ ). The computational algorithm aligns protein oligomers along a shared symmetry axis ( $C_2$ ), and generates an ensemble of protein-protein interfaces by translating and rotating about the shared symmetry axis. In order to reach fractional dimensional topology, we introduced controlled stochasticity along the shared symmetry axis by accepting fusion-domain-added sequences that could adopt multiple energetically favorable members of our protein-protein interface ensemble. We developed a coarse-grained simulation that allowed us to analyze emergent topological patterns based on Boltzmann-weighted probabilities of the stochastic rotations about the shared symmetry axis. A comparison of our simulations to cryo-tomography data of the protein assembly show excellent agreement with the number of binding partners and fractal dimension. We show that co-assembly of enzyme pathway members increases pathway efficiency, but not significantly more than the control "globular" assembly with 10xGSS extended linkers. However, the fractal performed significantly better than the control in sequestration of an IgG antibody, implicating channel porosity size as a key design consideration for future pathway co-assembly. The computational programs and simulations shared in this dissertation should enable the bottom-up design of symmetry-driven self-assembly, and the prediction of emergent topological properties as a result. Collectively, these studies should enable future efforts aimed at uncovering the fundamental design principles for functional metalloproteins and responsive supramolecular assembly with chemistry and functionality beyond those found in Nature.

## Acknowledgements

I would like to first thank my advisor, Dr. Sagar Khare, whose mentorship, patience, support, jokes, insight, and friendship have allowed me to carry out this work. I would also like to thank the other members of my committee, Drs. Spencer Knapp, Vikas Nanda, and Enver Izgu for their mentorship, collaborations, suggestions, and feedback. To Drs. Stephen Burley, Gail Ferstandig-Arnold, and the rest of the Institute for Quantitative Biomedicine at Rutgers, thank you for giving me a chance to excel and spread my wings. To Drs. Rudi Fasan, Nazomi Ando, and Wei Dai I would like to thank you for our collaborations and wonderful discussions.

I have had the great fortune to work with so many amazing people during my time here at Rutgers. To Khare lab members old and new (Srinivas, Manasi, Lu, Kenny, Kristin, Nancy, Brahm, Elliot, Dmitri, Denzel, Maria, Marium, BIMOD, Joey, Maria, Marissa, Ashley, Changpeng, Zhuofan, Lingjun, and many more), you are my family and I thank you for the many wonderful adventures we have gone on together. To my dearest friends who suffered me as a roommate during my time as a graduate student, Andrew Herschman and Yotam Cohen, I sincerely thank you. To my inspirations to continue higher education: Jason Karpinski, Eric Goll, Bill Boyke, and Britt Carlson, without you to fan the flame of science, none of this would be possible. A heartfelt thank you to my family, immediate and extended, for all you have done to help me grow into the person I have become today. Finally, thank you to my better half, Rachel Levitt, for believing in me and motivating me when I sometimes couldn't, you are my bashert.

## **Dedication**

To the pillars I stand upon,  
may I do the same.

# Table of Contents

<b>ABSTRACT OF THE DISSERTATION .....</b>	<b>ii</b>
<b>Acknowledgements .....</b>	<b>v</b>
<b>Dedication .....</b>	<b>vi</b>
<b>Table of Contents .....</b>	<b>vii</b>
<b>List of Tables .....</b>	<b>xii</b>
<b>List of Illustrations.....</b>	<b>xiv</b>
<b>Chapter 1:     <b>Introduction .....</b></b>	<b>1</b>
1.1     Introduction to Multinuclear Metalloprotein Design .....	1
1.2     Introduction for Supramolecular Protein Assembly Design.....	13
1.3     References.....	21
<b>Chapter 2:     <b>Computational Design of Multinuclear Metalloproteins Using</b></b>	
<b>Unnatural Amino Acids.....</b>	<b>24</b>
2.1     Preface.....	24
2.2     Summary .....	24
2.3     Introduction.....	25

2.4	Methods.....	28
2.5	References.....	40

**Chapter 3:     Benchmarking a computational design method for the incorporation  
of metal ion-binding sites at symmetric protein interfaces ..... 43**

3.1	Preface.....	43
3.2	Abstract.....	43
3.3	Introduction.....	44
3.4	Results.....	46
3.5	Discussion.....	57
3.6	Methods.....	60
3.7	References.....	65
3.8	Supplementary Material.....	73

**Chapter 4:     Enzyme stabilization via computationally guided protein stapling.. 83**

4.1	Preface.....	83
4.2	Abstract.....	83
4.3	Introduction.....	84
4.4	Results.....	86
4.5	Discussion.....	98

4.6	References.....	101
4.7	Supplementary Information.....	108
4.8	Supplemental References .....	141
<b>Chapter 5:</b>	<b>A site-selective protein modification for N2S* metal chelation .....</b>	<b>143</b>
5.1	Preface.....	143
5.2	Abstract .....	143
5.3	Introduction.....	144
5.4	Results and Discussion.....	145
5.5	Conclusion .....	156
5.6	Supplemental Information.....	157
<b>Chapter 6:</b>	<b>Design and Evolution of a Macrocyclic Peptide Inhibitor of the Sonic Hedgehog/Patched Interaction .....</b>	<b>188</b>
6.1	Preface.....	188
6.2	Abstract .....	188
6.3	Introduction.....	189
6.4	Results and Discussion.....	192
6.5	Conclusion .....	209
6.6	References.....	211

6.7	Experimental Procedures .....	220
6.8	Supporting Information .....	230

**Chapter 7: Structures of the peptide-modifying radical SAM enzyme SuiB  
elucidate the basis of substrate recognition ..... 259**

7.1	Preface.....	259
7.2	Abstract .....	259
7.3	Introduction.....	260
7.4	Results .....	264
7.5	Discussion .....	275
7.6	References.....	278
7.7	SI Results and Discussion (cont.).....	284
7.8	SI Materials and Methods .....	285

**Chapter 8: Stimulus-responsive self-assembly of protein-based fractals by  
computational design ..... 308**

8.1	Preface.....	308
8.2	Abstract .....	308
8.3	Main .....	309
8.4	Results.....	312
8.5	Discussion .....	330

8.6	References.....	331
8.7	Supplemental Information.....	337

## List of Tables

Table 1-3. Table S1.....	79
Table 2-3. S1. Continued. ....	80
Table 3-3. Table S2.....	81
Table 4-3. S2. Continued. ....	82
Table 5-4. Computational and experimental values for Mb(H64V,V64A) and its variants .....	89
Table 6-4. Table 2. Catalytic activity and selectivity of Mb(H64V,V68A) and its stapled variants for cyclopropanation of styrene (1a) and pentafluorostyrene (1b) with ethyl 2- diazoo-acetate (2) in the absence and in the presence of organic cosolvents .....	96
Table 7-4. Table S1. Oligonucleotide sequences.....	139
Table 8-4. Table S2. Rosetta scores of designs sMb1-9 after energy minimization and rotameric sampling.....	141
Table 9-5. Table 1. EPR hyperfine couplings and g-factors .....	153
Table 10-5 Table 2: Crystallographic data for [Zn(BMIE-CPG2-S203C)] <sup>2+</sup> .....	155
Table 11-6. Table S1. Oligonucleotide Sequences .....	246
Table 12-6. S2 MS data and retention times for linear and cyclic L2 mimics. ....	247
Table 13-7. Table S1. Crystallographic data processing and refinement statistics for SuiB structures. ....	297
Table 14-7. Table S2. Missing residues for each structure.....	297
Table 15-9. Supplementary Table 1. List of substitutions and reasons for the various AtzA and AtzC designs.....	370
Table 16-8. Supplementary Table 2. Curve fitting data for Supplementary Figure 16. .	386

Table 17-8. Supplementary Table 3. Comparison of the different AtzA and AtzC ratio components with their fractal dimensions ( $D_f$ ) and $\lambda$ . .....	402
--	-----

## List of Illustrations

Figure 1-1. Figure 1. Outline of multi-nuclear metalloprotein precursor design.....	3
Figure 2-1.....	14
Figure 3-2. Figure 1. Several target cofactors that this method was intended to implement using scaffolds of various symmetries.....	27
Figure 4-2. Figure 2. Method overview, incorporation of a $\text{Co}_4\text{O}_4(\text{Ac})_2(\text{bipyridine})_4$ cofactor with noncanonical amino acids into a D2 symmetric scaffold. ....	28
Figure 5-2. Figure 3. SyPRIS design flowchart.....	30
Figure 6-2. Figure 4. Alignment of backbone atoms.....	33
Figure 7-2. Figure 5. SyPRIS design output. ....	37
Figure 8-3. Figure 1. The computational workflow of the SyPRIS algorithm. ....	48
Figure 9-3. Figure 2. Native SyPRIS match results for a benchmark of 67 native metal chelating homo-oligomeric proteins. ....	49
Figure 10-3. Figure 3. Comparison of native and non-native matches with SyPRIS score and Rosetta residue energy. ....	52
Figure 11-3. Figure 4. Cross-match results for symmetric CMPs composed of three histidine residues into trimeric scaffolds at native metal-binding histidine locations. ....	57
Figure 12-3. Figure 5. SyPRIS-related terms and parameters. ....	62
Figure 13-4. Figure 1. Computational design approach. ....	87
Figure 14-4. Figure 2. Characterization of stapled Mb(H64V,V68A) variants (sMb variants).....	91
Figure 15-4. Figure S1. Near-attack conformation (NAC) analysis.....	118
Figure 16-4. Figure S2. NAC analysis for additional sMb variants. ....	119

Figure 17-4. Figure S3. Interaction between O2beY36 and Phe106 in sMb2. ....	120
Figure 18-4. Figure S4. Visible-range electronic absorption spectra for Mb(H64V,V68A) and representative sMb variants in the ferric form. ....	121
Figure 19-4. Figure S4. (cont.).....	122
Figure 20-4. Figure S5. SDS-PAGE gel of remaining sMb variants not included in Figure 2A.....	123
Figure 21-4. Figure S6. MALDI-TOF MS spectrum of additional sMb variants not included in Figure 2B.....	124
Figure 22-4. Figure S6 (cont.).....	125
Figure 23-4. Figure S7. Thermal denaturation curves for wild-type Mb, Mb(H64V,V68A) and sMb variants. ....	126
Figure 24-4. Figure S7 (cont.).....	127
Figure 25-4. Figure S7 (cont.).....	128
Figure 26-4. Figure S7 (cont.).....	129
Figure 27-4. Figure S7 (cont.).....	130
Figure 28-4. Figure S7 (cont.).....	131
Figure 29-4. Figure S7 (cont.).....	132
Figure 30-4. Figure S8. Overlay of near-UV circular dichroism spectra corresponding to Mb(H64V,V68A) and stapled variants sMb5, sMb10, and sMb13.....	132
Figure 31-4. Figure S9. Chemical denaturation curves for Mb(H64V,V68A) and stapled variants sMb5, sMb10, and sMb13 in the presence of guanidium chloride. ....	133
Figure 32-4. Figure S9 (cont.).....	134

Figure 33-4. Figure S10. Catalytic activity (top graph) and diastereo- and enantioselectivity (bottom graph) of Mb(H64V,V68A) and stapled variants in styrene cyclopropanation reaction in the presence of 30% v/v methanol (MeOH). ....	135
Figure 34-4. Figure S11. Catalytic activity (top graph) and enantioselectivity (bottom graph) of Mb(H64V,V68A) and stapled variants in styrene cyclopropanation reactions in the presence of 30% and 45% v/v DMF. ....	136
Figure 35-4. Figure S12. Catalytic activity (top graph) and enantioselectivity (bottom graph) of Mb(H64V,V68A) and stapled variants in styrene cyclopropanation reactions in the presence of 30% and 45% v/v DMSO. ....	137
Figure 36-4. Figure S13. SFC analysis of cyclopropanation products from reaction with pentafluorostyrene and EDA.....	138
Figure 37-5. Figure 1. BMIE characterization with small molecule BMIE-thiol adducts. ....	148
Figure 38-5. Figure 2. CPG2 modification and site specific cysteine labeling with BMIE. ....	150
Figure 39-5. Figure 3. EPR characterization of binary and ternary BMIE:Cu complexes. ....	153
Figure 40-5. Figure 4. Electron density at the Cys203-BMIE-Zn <sup>2+</sup> site. The 2F <sub>o</sub> -F <sub>c</sub> 3.0 Å from the sidechain atoms of residue 203 (both subunits) is shown contoured at 0.9 σ. The coordination of atoms with the Zn <sup>2+</sup> ion is shown as red dashed lines.....	155
Figure 41-5. Figure S1. Mass spec date of BMIE-cys, expected MW 424 da.....	172
Figure 42-5. Figure S2. H <sup>1</sup> -NMR and C <sup>13</sup> -NMR of BMIE-cys.....	173
Figure 43-5. Figure S3. Ellman's Assay: BMIE-l-cys conjugation over time. ....	174

Figure 44-5. Figure S4. UV/Vis spectra of BMIE-I-cys and different divalent cations (Co, Ni, Cu, and Zn). .....	175
Figure 45-5. Figure S5. Full NMR spectra (stacked graph) of the complex of BMIE-Cys with zinc chloride.....	176
Figure 46-5. Figure S6. Full NMR spectra (stacked graph) of the complex of BMIE-Cys with zinc triflate. ....	177
Figure 47-5. Figure S7. Full NMR spectra (stacked graph) of the complex of BMIE-Cys with tetrakis acetonitrile copper (I) tetrafluoroborate. ....	178
Figure 48-5. Figure S8. Crystal structure of copper chloride in complex with the BMIE-TC ligand. ....	179
Figure 49-5. Figure S9. Time series Ellman's assay of 20uM protein in 100mM NaPO <sub>4</sub> , 1mM EDTA, pH 8.0 with varying concentrations of BMIE. ....	180
Figure 50-5. Figure S10. Two-pulse echo-detected EPR spectra for a small-molecule complex BMIE:Cu (2:1) in MOPS buffer at pH 7.5 (top) and a binary CPG2-BMIE:Cu cluster in Tris buffer at pH 8.5 (bottom), measured at 22 K.....	181
Figure 51-5. Figure S11. Two-pulse ESEEM decays (A, B) and their cosine Fourier-transform (FT) spectra (C, D) for a small-molecule complex BMIE:Cu (2:1) in MOPS buffer at pH 7.5 and binary/ternary CPG2-BMIE:Cu clusters in Tris and MOPS buffers at pH 7.5-8.5, as labeled. ....	182
Figure 52-5. Figure S12. Three-pulse ESEEM decays for (A) a small-molecule complex BMIE:Cu (2:1) in MOPS buffer at pH 7.5 and (B) binary/ternary CPG2-BMIE:Cu clusters in Tris and MOPS buffers at pH 7.5-9.6, as labeled.....	183

Figure 53-5. Figure S13. Cosine FT spectra of three-pulse ESEEM experiment for a small-molecule complex BMIE:Cu (2:1) in MOPS buffer at pH 7.5. ....	184
Figure 54-5. Figure S14. Cosine FT spectra of three-pulse ESEEM experiment for a binary complex CPG2-BMIE:Cu in Tris buffer at pH 8.5.....	185
Figure 55-5. Figure S15. Davies ENDOR spectra for a small-molecule complex BMIE:Cu (2:1) in MOPS buffer at pH 7.5 and several binary/ternary CPG2-BMIE:Cu clusters in Tris and MOPS buffers at pH 7.5-8.5, as labeled, measured at 22 K and magnetic field 285-287 mT corresponding to the $g_{\parallel}$ field orientations. ....	186
Figure 56-5. Figure S16. Davies ENDOR spectra for two small-molecule complexes BMIE:Cu (2:1) and PHEN:Cu (2:1) and a ternary CPG2-BMIE:Cu:PHEN cluster. Samples in in MOPS buffer at pH 7.5, as labeled, measured at 22 K and magnetic fields (A) 285 mT and (B) 337-342 mT corresponding to the $g_{\parallel}$ and $g_{\perp}$ field orientations, respectively. Non-selective microwave pulses ( $\tau$ pulse = 32 ns) were applied in order to suppress contributions from weakly-coupled $^1\text{H}$ protons. Vertical lines mark the peak positions from the BMIE- and PHEN-derived $^{14}\text{N}$ nitrogens directly-coordinated to $\text{Cu}^{2+}$ . On both panels the bottom thin line traces show that the spectra in CPG2-BMIE:Cu:PHEN (green) can be roughly represented as a 1:1 sum of the spectra in BMIE:Cu (2:1) (dark blue) and PHEN:Cu (2:1) (light blue) at both $g_{\parallel}$ and $g_{\perp}$ field orientations.....	187
Figure 57-6. Figure 1. Hedgehog signaling pathway.....	190
Figure 58-6. Figure 2. Macrocyclic HHIP L2 loop mimic. ....	194
Figure 59-6. Figure 3. The shh-binding affinity of linear and macrocyclic L2 mimics. ....	196
Figure 60-6. Figure 4. Overview of strategy for the evolution of macrocyclic peptides.....	197

Figure 61-6. Figure 5. Circular dichroism (CD) .....	202
Figure 62-6. Scheme 1. Synthesis of Macrocyclic Peptides:.....	204
Figure 63-6. Figure 6. HL2-m5-induced suppression of Hh pathway signaling. ....	206
Figure 64-6. Figure 7. Hh analog selectivity of linear and cyclic L2 mimics. ....	208
Figure 65-6 Figure S1. Close-up view of HHIP L2 loop interaction with Shh (pdb 3HO5). .....	236
Figure 66-6. Figure S2. Model of evolved macrocyclic peptide HL2-m5 in complex with Shh. ....	237
Figure 67-6. Figure S3. Relative Shh binding activity for representative hits from the single-site sitesaturation libraries.....	238
Figure 68-6. Figure S4. Relative Shh binding activity for representative hits from the multi-site recombinant libraries. ....	238
Figure 69-6. Figure S5. SDS-PAGE gel of recombinantly expressed GST-Shh, GST-Ihh, and GST-Dhh after purification by Ni-affinity chromatography.....	239
Figure 70-6. Figure S6. Thiol-induced intein cleavage reactions. ....	239
Figure 71-6. Figure S7. MALDI-TOF MS spectra corresponding to purified FLAG- tagged linear and cyclic L2 mimics obtained via recombinant expression. ....	240
Figure 72-6. Figure S8. Analytical HPLC chromatogram (A) and ESI-MS spectra in positive (B) and negative mode (C) corresponding to synthetic HL2-m5. Y* = alkylated O <sup>2</sup> beY. ....	242
Figure 73-6. Figure S9. Analytical HPLC chromatogram (A) and ESI-MS spectra in positive (B) and negative mode (C) corresponding to synthetic HL2-m1. Y* = alkylated O <sup>2</sup> beY. ....	244

Figure 74-6. Figure S10. Inhibition curve corresponding to HL2-m5 induced inhibition of FLAG-HL2- m5 binding to plate-immobilized GST-Shh. ....	244
Figure 75-6. Figure S11. Dose-response curves for direct binding of FLAG-HL2-m5 to plateimmobilized GST-Shh, GST-Ihh, or GST-Dhh as determined using the colorimetric assay with HRP-conjugated anti-FLAG antibody. ....	245
Figure 76-6. Figure S12. Proteolytic stability of linear and cyclic L2 mimics.....	245
Figure 77-7. Figure 1. The <i>sui</i> gene cluster and the reaction catalyzed by SuiB.....	262
Figure 78-7. Figure 2. SuiA recognition in the active site is dominated by interactions of the leader sequence with the bridging region. ....	268
Figure 79-7. Figure 3. Substrate binding leads to coordinated loop movements. ....	271
Figure 80-7. Figure 4. Binding of the SuiA leader sequence supports positioning of the core sequence in the active site of SuiB.....	272
Figure 81-7. Figure S1. SAM binding and cleavage in the SuiB active site. ....	298
Figure 82-7. Figure S2. Sequence alignment of SuiB with anSMEcpe.....	299
Figure 83-7. Figure S3. Comparison of the RRE domain in SuiB with those previously characterized by Xray crystallography. ....	300
Figure 84-7. Figure S4. Sequence alignment of SuiB homologs.....	301
Figure 85-7. Figure S5. Leader peptide binding site of SuiB. ....	302
Figure 86-7. Figure S6. Conformational changes in SuiB upon binding of substrate SuiA. ....	303
Figure 87-7. Figure S7. Energy landscape of the cyclized SuiA peptide in the SuiB active site. ....	304

Figure 88-7. Figure S8. Rosetta simulations yielded four sets of distinct low energy conformations for the cyclized peptide when SAM was replaced with 5'-dA.....	305
Figure 89-7. Figure S9. Structural comparison with anSMEcpe. ....	306
Figure 90-7. Figure S10. The simulated location of the Lys-to-Trp crosslink overlays well with the H-atom abstraction sites of (A) anSMEcpe (20) and (B) RlmN (21). ....	307
Figure 91-8. Fig. 1: Multiscale computational design approach for fractal assembly design. ....	311
Figure 92-8. Fig. 2: Assembly formation, dissolution and inhibition in vitro. ....	317
Figure 93-8. Fig. 3: Assembly formation and characterization with helium ion microscopy, AFM and transmission electron microscopy. All methods reveal fractal-like topologies on a surface.....	321
Figure 94-8. Fig. 4: Assembly characterization with cryo-ET. ....	323
Figure 95-8. Fig. 5: Fractal assemblies capture and release greater amounts of cargo compared to globular assemblies. ....	328
Figure 96-8. Supplementary Figure 1. Scheme for designing arboreal fractal morphologies.....	367
Figure 97-8. Supplementary Figure 2. Flowchart of interface and linker design method. ....	368
Figure 98-8. Supplementary Figure 3. Design considerations for selecting substitutions and atomic interactions responsible for orientations of components during simulation.	369
Figure 99-8. Supplementary Figure 4. Illustration of the box-counting algorithm. ....	371
Figure 100-8. Supplementary Figure 5. Computational parameter sweep of $kT$ (major y-axis), $P_{\text{term}}$ (minor y-axis), and $C_{\text{frac}}$ (minor x-axis). ....	372

Figure 101-8. Supplementary Figure 6. Representative simulated fractal images .....	373
Figure 102-8. Supplementary Figure 7. Phosphorylation of SH2 peptide AtzA fusion (pY-AtzA) by Src kinase. ....	374
Figure 103-8. Supplementary Figure 8. Experimental selection process for pY-AtzA and AtzC-SH2.....	375
Figure 104-8. Supplementary Figure 9. Experimental selection of AtzA, AtzC subunits for characterization. ....	376
Figure 105-8. Supplementary Figure 10. Biolayer interferometry (BLI) binding profiles of AtzC wildtype SH2 fusion (AtzC-wtSH2) and AtzC superbinder SH2 fusion (AtzC- SH2) to phosphorylated SH2 binding peptide AtzA fusion (pY-AtzA). ....	377
Figure 106-8. Supplementary Figure 11a. Sequence alignment of AtzC-SH2 designs AtzCM0-AtzCM5. ....	378
Figure 107-8. Supplementary Figure 11b. Sequence alignment of AtzC-SH2 designs AtzCM0-AtzCM5 (con't).....	379
Figure 108-8. Supplementary Figure 12. Sequence alignment of pY-AtzA designs.....	380
Figure 109-8. Supplementary Figure 13. Visual assembly turbidity.....	381
Figure 110-8. Supplementary Figure 14. Inhibition of assembly at 0.66 $\mu$ M AtzC-SH2, 1 $\mu$ M pY-AtzA, 0-6 $\mu$ M SH2-DhaA.....	382
Figure 111-8. Supplementary Figure 15. Inhibition of assembly at 2 $\mu$ M AtzC-SH2, 3 $\mu$ M pY-AtzA with 0-15 $\mu$ M inhibitor.....	383
Figure 112-8. Supplementary Figure 16. Inhibition of assembly at 2 $\mu$ M AtzC-SH2, 3 $\mu$ M pY-AtzA with 0-15 $\mu$ M inhibitor (con't). ....	384

Figure 113-8. Supplementary Figure 17. Rate of assembly formation is dependent on ATP concentration. ....	385
Figure 114-8. Supplementary Figure 18. Bright-field view of the assembly growing after the addition of Src kinase.....	387
Figure 115-8. Supplementary Figure 19. Average size of particle formed by pY-AtzA and wild type AtzC-SH2.....	388
Figure 116-8. Supplementary Figure 20. Average size of particle formed by pY-AtzA and super-binder AtzC-SH2. ....	389
Figure 117-8. Supplementary Figure 21. Helium ion microscopy (HIM) depict fractal-like assembly with increasing AtzA concentrations. ....	390
Figure 118-8. Supplementary Figure 22. Atomic Force Microscopy (AFM) images show fractal-like structures, fern-like, and petal-like structures, similar to HIM. ....	391
Figure 119-8. Supplementary Figure 23. Helium ion microscopy (HIM) buffer and non-phosphorylated controls preclude salt precipitation. ....	392
Figure 120-8. Supplementary Figure 24. Helium ion microscopy comparison of fractal assembly and globular assembly.....	393
Figure 121-8. Supplementary Figure 25. Transmission Electron Microscopy (TEM) depicts fractal-like assemblies in the phosphorylated samples while the non-phosphorylated samples depict individual proteins. ....	394
Figure 122-8. Supplementary Figure 26. Comparison of the fractal assembly CryoEM tomograms and the extended linker globular assemblies. ....	395
Figure 123-8. Supplementary Figure 27. Analysis of the fractal assembly CryoEM tomograms and the extended linker globular assemblies. ....	396

Figure 124-8. Supplementary Figure 28. Spatial angle comparison of Cryo-EM structure to simulation.....	397
Figure 125-8. Supplementary Figure 29. Isosurface views of the assembly tomograms, from large to small. ....	398
Figure 126-8. Supplementary Figure 30. Fluorescence microscopy and bright-field images of the 4-component assembly (AtzAM1, AtzCM1, ProteinA-SH2, and antibody, along with extended linker versions of AtzA and AtzC). ....	399
Figure 127-8. Supplementary Figure 31. Helium ion microscopy (HIM) images depict fractal-like assembly with 3 $\mu$ M AtzAM1, 1 $\mu$ M AtzBSH2, 1 $\mu$ M AtzCM1 final protein concentrations. ....	400
Figure 128-8. Supplementary Figure 32. Helium ion microscopy (HIM) images depict fractal-like assembly with 3 $\mu$ M AtzAM1, 1 $\mu$ M AtzBSH2, 2 $\mu$ M AtzCM1 final concentrations. ....	401
Figure 129-8. Supplementary Figure 33. DLS and SDS PAGE confirm AtzBSH2 incorporation into the 3-component assembly.....	403
Figure 130-8. Supplementary Figure 34. Fluorescence microscopy and bright-field images of the 3-component assembly confirm incorporation of AtzBSH2 into assembly while bright-field images confirm the fractal-like nature of the 2-component assembly. ....	404
Figure 131-8. Supplementary Figure 35. AtzBSH2 incorporation to construct a three-enzyme assembly. ....	405

Figure 132-8. Supplementary Figure 36. Phase contrast micrographs of the Basotect® polymer foam with and without assemblies for the AtzAM1, AtzBSH2, and AtzCM1 components. ....	406
Figure 133-8. Supplementary Figure 37. The fractal-like assemblies (Reg-Assembly) and the extended linker globular assemblies (ExtLinker-Assembly) enzymatic conversion of atrazine to cyanuric acid demonstrates no enzymatic benefit of a globular assembly. ..	407

## Chapter 1: Introduction

Despite advances in the development of *de novo* proteins design, metalloenzymes and supramolecular assembly present a fundamental challenge and remain largely unexplored areas in enzyme design and engineering. To expand the field of protein design, we have developed and implemented design strategies for the incorporation of multinuclear metal clusters (atomic scale) and the self-assembly of symmetric protein molecules (macromolecular scale). The approaches herein utilize the inherent benefits of protein symmetry to achieve these goals.

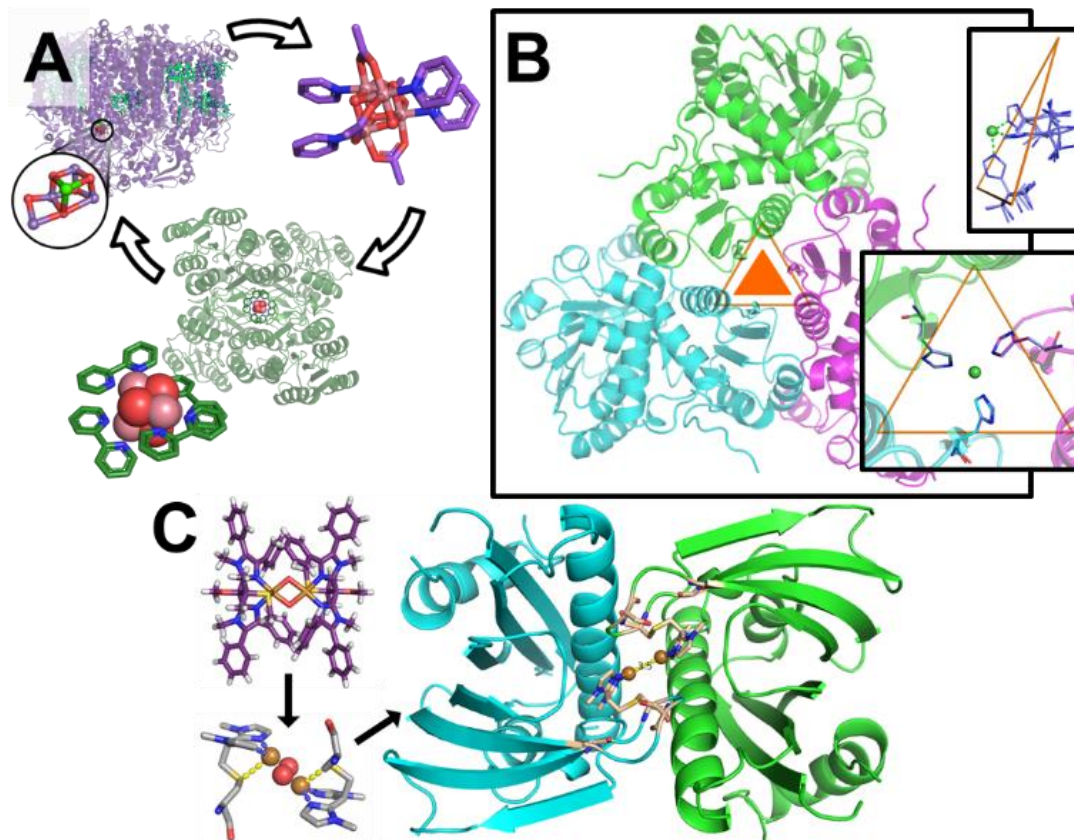
### 1.1 Introduction to Multinuclear Metalloprotein Design

Billions of years of evolution under selective pressure have led to the development of structurally sophisticated protein molecules that carry out most of life's biological functions<sup>1</sup>. Among these protein molecules, multinuclear metalloproteins (MNMPs) perform the most exquisite chemistry, including water oxidation, nitrogen fixation, and hydrogenation. While efforts have been made to mimic natural MNMPs with small-molecule coordination complexes, the mimics fail to compete with their natural counterparts. For example, the oxygen-evolving complex in photosystem II is 25,000 times faster and can survive a million-fold more turnovers than the coordination complex  $\text{Co}_4\text{O}_4(\text{pyridine})_4(\text{acetate})_4$ , a small molecule mimic<sup>2</sup>. It is well established that protein scaffolding can provide a protective environment for multinuclear metallocusters to safeguard from oxidative damage and complex decay. Despite advances in technology and the advent of computational software like Rosetta<sup>3</sup>, which have greatly improved protein design accuracy, the design of metalloproteins, let alone MNMPs, remains a

considerable challenge<sup>4</sup>. The design of a metal site requires the accurate positioning of multiple coordinating amino-acids, first shell interactions to stabilize the coordinating sidechains, and polarizing second shell interactions to promote metal affinity and tune redox potential<sup>5</sup>. The combinatorial complexity of positioning a myriad of amino acids coupled with the metal ions inherent ability to adopt multiple coordination geometries leads to the wide Nature-human design gap present in the design of metalloproteins<sup>4</sup>.

Two methods have greatly improved metal design accuracy: polydenticity<sup>6,7,8</sup> and symmetry<sup>7,9</sup>. Polydenticity can directly reduce the number of amino acid substitutions required to design a coordination complex. Outside of the canonical amino acids (Asp and Glu), reliable polydentate coordination has been achieved using synthetic cofactor coordination<sup>10</sup>, chemical modification<sup>8</sup>, or recombinant incorporation of non-canonical amino acids (ncAA)<sup>6,7</sup>. Symmetry also offers a method for reducing the design complexity, albeit indirectly. When the strategies combine, the introduction of a single bidentate ncAA in a spacious interface of a  $C_3$  symmetric protein can create a metal-binding site complete with six coordinating atoms<sup>7</sup>. A similar site using traditional design methods would typically need to introduce six natural amino acids either in an available pocket—or more invasively, within a newly designed pocket<sup>11</sup>. If we are going to tackle future challenges, like the development of clean and renewable energy from water, we must understand the fundamental principles that govern the accurate design of multinuclear metalloproteins (MNMPs). In section 1.1. I will define, develop, and use a methodological design framework to understand the first principles for designing multinuclear metalloproteins. Additionally, I will develop computational and molecular

tools for making functional metalloproteins. Finally, I will discuss expanding the biomolecular toolbox with new derivatized amino acids through chemical conjugation.



**Figure 1-1.** Figure 1. Outline of multi-nuclear metalloprotein precursor design.

(A) Natural MNMPs inspire the development of small molecule mimics. To bridge the gap, we introduce an MNMP precursor—a protein modified to house a novel metal cluster. (B) SyPRIS protocol overview (chapter 2&3): Select the desired metal coordination or "probe" (top right), match the symmetry of the probe to a library of oligomeric proteins with the same symmetry. Align probe and protein symmetry axes (orange triangle) and locate backbone positions for the desired coordination complex. (C) Method workflow example: design of a dinuclear copper protein using BMIE (chapter 5).

### 1.1.1 Statement of the Problem

To close the cycle and bridge the nature-human design gap, we must go beyond the design of small molecule mimics and introduce the idea of an MNMP precursor—a *de novo* metalloprotein that incorporates desired multinuclear coordination complexes into protein scaffolds. The accurate bottom-up design of MNMP precursors, with precise metal coordination in multiple protein scaffolds, should enable the development of functional metalloenzymes that can be further tuned to rival nature. Richard Feynman put it best, "What I cannot create, I do not understand." This dissertation attempts to answer the following questions: 1) Can we develop a general design strategy for the accurate positioning of amino-acid constellations to incorporate novel metal binding sites? 2) How can we contribute to the computational and molecular toolbox that already exists? 3) Using the design strategy and generated tools, can we design a protein to house a non-native multi-nuclear metal site from the first principles?

### 1.1.2 General Design Heuristics

Catalytic efficiency, defined by TOF and TON, of small organometallic molecules, often fails to compete with enzymes capable of the same function. Protein environments are highly selective and can protect from oxidative damage, unlike their organometallic mimics. Despite the attractive features of proteins, organometallic mimics remain the industry gold standard for catalysis. They are cheaper and easier to make, tolerant of high temperatures, and offer a broad catalytic spectrum. The ligand environment around the metal center can be fine-tuned systematically to promote stereoselectivity or enhance yield for the desired reaction. Contrarily, metalloenzymes with a desirable metal-site are

difficult to tune for non-native reactions—almost every amino acid plays a role in coordination or folding the structure. We wanted to create a catalyst with the versatility to be evolved for target reactions but with the protection and efficiency granted by being within a protein environment. Furthermore, the idea of using protein oligomers as large symmetric "ligands" was born. Over 1 million small molecule crystal structures are available today on the Cambridge Crystallographic Data Center; many are symmetric metal-complexes—offering a large set of starting structures for design.

The first step is to identify an organometallic complex that performs the target reaction. We then want to pair the symmetry of coordination with oligomeric proteins sharing that symmetry. Oligomeric interfaces are often porous and less hydrophobic than the folded cores of the symmetric monomers—creating the ideal design environment. Additionally, using the symmetric protein interface enhances tunability as each amino acid substitution is copied to all chains symmetrically. The same set of amino-acid derivative will be symmetrically coordinated by the eventual complex. After locating a suitable scaffold, we will design the surrounding environment to incorporate the cluster—creating an MNMP precursor.

If we can harness the versatility that organometallic complexes offer and combine the advantages of a protein matrix by incorporating them within a protein scaffold, then we will enable the general design of tunable biocompatible catalysts.

### 1.1.3 Matching the Desired Coordination with Symmetric Protein Interfaces

We propose to repurpose known symmetric interfaces of oligomeric proteins for the coordination of metal ions using chemically modified or dative polydentate interactions. To achieve this, we developed a program called Symmetric Protein Recursive Ion-cofactor Sampler (SyPRIS)<sup>12</sup>. SyPRIS takes a model of the desired coordination complex with symmetrically bound free amino acid side-chains as ligands, generates an inverse rotamer ensemble, and probes the symmetric interface of a library of homo-oligomeric protein interfaces for compatible backbone positions. The model of the complex with the inverse rotamer library has been named a chelant model probe (CMP). The method by which SyPRIS matches is covered in Chapter 2 and Chapter 3. SyPRIS was benchmarked using 67 natural symmetric proteins ( $C_2$  and  $C_3$ ) with metal coordination sites along the symmetric interface and was able to recapitulate all 67 natural coordination geometries<sup>13</sup>. In a cross-match study, SyPRIS predicted >1000 different novel metal coordination sites with native-like match scores showing the potential for SyPRIS to produce matches for novel non-native metal coordination complexes<sup>13</sup>. The details of the benchmark are covered in Chapter 3. SyPRIS is extendable to all dative and chemical modifications so long as they coordinate symmetrically. The success of this computational tool highlights the usability of incorporating novel metal complexes into symmetric protein interfaces and provides a platform to incorporate metal-binding using ncAA and chemical modifications.

#### 1.1.4 Modeling Non-canonical amino acids

When designing MNMPs, each asymmetric metal ion requires coordinated at one or more sites. One way to reduce the number of unique amino-acids is to introduce amino-acid derivatives with multiple coordinating atoms or polydententicity. Fewer amino acids greatly reduce the search space and increase the likelihood of finding symmetric backbones suitable for the desired coordination. Physics-based computational tools, like Rosetta, are uniquely equipped to work with ncAAs and other non-protein chemical cofactors or modifications.

In Chapter 4, we discuss in detail how we successfully modeled the ncAA O-2-bromoethyl tyrosine (O2beY) within Rosetta and produced several cyclic-myoglobin structures with increased thermostability. When a free thiol, i.e., cysteine, is in proximity with O2beY, O2beY will undergo nucleophilic substitution and release free bromine. In this study, we modeled a conjugated O2beY-Cys and located within the crystal structure of myoglobin (PDB: 1JP9) two backbone positions that would be capable of forming the O2beY-Cys crosslink. This approach produced 9 single-crosslinked designs to be tested by the Fasan lab. Their results showed that 4/9 designs produced a cyclized form of myoglobin. Interestingly, when we made a second model of the O2beY-Cys in a near attack confirmation (NAC) and attempted the same matching protocol, only the 4 experimentally verified designs were found in the match results. This result details the atomic accuracy necessary for forming successful crosslinks and highlights the need for computational-based methods for generating structures with specific geometric requirements. Although this molecule is not symmetric and does not bind metal ions, this

matching method is analogous to SyPRIS and is a testament to our ability to successfully model ncAAs.

#### 1.1.5 Expanding the Molecular Toolbox

Despite the growing (>200) ncAAs that have been synthesized and introduced recombinantly, only 2 exhibit high-affinity polydentate metal binding. The metalloprotein design field is starved for new tools to localize metal ions. Chemical conjugation of existing natural amino acids offers a much wider variety of chelating options, but all are bulkier than their recombinantly introduced counterparts. Furthermore, the conjugated products often possess flexible sidechains with multiple rotatable bonds—increasing complexity while decreasing predictability in design.

To expand the biomolecular toolbox, we, in collaboration with the Knapp lab, designed bis(1-methyl-1H-imidazole-2-yl)ethene or BMIE, a compact high-affinity metal-coordination ligand capable of thiol conjugation. The bidentate 6-member bite is bridged by an  $\alpha,\beta$ -unsaturated vinylidene, a substituent shown to conjugate free thiols i.e. cysteine. Compared to the two recombinantly introduced ncAAs, a BMIE-cys adduct is similar in size and possesses one additional rotatable bond—making it the smallest chemical modification with a high-affinity for metal ions known.

General thiol reactivity was confirmed by conjugate additions of various small-molecule thiols (l-cys, boc-cys, thiocresol, and glutathione). BMIE-thiol adducts were shown to coordinate multiple divalent metal ions (Co, Ni, Cu, and Zn) in both bidentate ( $N_2$ ) and tridentate ( $N_2S^*$ ) coordination geometries. To test viability as a bioconjugate, we

modified a carboxypeptidase G2 variant (S203C), containing a single cysteine. Ellman's reagent, which measures the amount of free thiol, was used to determine more than 90% formation of the modified protein. ESI-MS confirmed that conjugation was site-selective for the single surface exposed Cys. EPR characterization of the modified protein and  $\text{CuCl}_2$  demonstrated 1:1 BMIE- $\text{Cu}^{2+}$  coordination with symmetric tetragonal geometry under physiological conditions. A highly sought feature of chemical modifications for metal binding is to allow for vacant or labile metal coordination sites about the metal center. Labile positions enable substrate coordination and often lead to catalytic function. The copper-bound modified protein was shown to accept a host of counter-ligands ( $\text{H}_2\text{O}/\text{HO}^-$ , tris, and phenanthroline) at the remaining equatorial positions opposite BMIE. Furthermore, x-ray derived structures of modified protein co-crystallized with  $\text{ZnCl}_2$  revealed a heterogeneous BMIE- $\text{Zn}^{2+}$  contact along the intermolecular crystallographic interface between a modified and an unmodified cysteine monomer. The crystal structure interface contact makes a provoking case for utilizing BMIE as a structural binding motif in future metal-directed interface design. These features (covered in Chapter 5), combined with ease of synthesis, make for an attractive metalloprotein design tool and should enable future catalytic and structural applications.

#### 1.1.6 Future of Symmetric MNMP Precursor Design

In the final phase of our design framework, we want to demonstrate that the developed computational and chemical tools are broadly useful for a host of design problems. After using SyPRIS to locate compatible backbones, we will symmetrically design the surrounding residues with Rosetta FastDesign to better accommodate the non-native

complex, i.e., removal of surrounding residues that sterically clash or interact unfavorably. These efforts will be the first MNMP precursors in the methodological framework and will be studied and perturbed extensively to uncover the fundamental principles that govern metalloproteins and catalysis. The efforts to realize these metalloproteins will be covered in future manuscripts.

#### 1.1.6.1 $\text{Cu}_2(\text{OH})_2$ Coordination

We sought to exploit BMIE's ability to bind copper ions to create a dinuclear  $\text{Cu}_2(\text{OH})_2$  complex for activated carbon oxidation reactions such as phenol and catechol oxidation. Cu-Cu distance in natural enzymes greatly influences the bound oxygen species ( $2\text{xOH}$ ,  $\text{HOOH}$ , and  $\text{O}_2$ ), which further dictates function. If we could display the fine control over Cu-Cu distance necessary to govern function, we would display the utility of our method and enable future MNMP precursor design efforts. To begin, we obtained the coordination geometry from a CCDC-derived BMIE analog coordinated to the same  $\text{Cu}_2(\text{OH})_2$  center. We developed a CMP from this structure using a post-conjugation BMIE-Cys "residue" and generated a BMIE-Cys inverse rotamer library. The CMP was used to probe a library of more than 10,000  $\text{C}_2$  protein structures from the PDB. We designed the output matches and selected seven designs of the coordination complex with varying ligand environments and surrounding protein scaffolds. Preliminary data suggest 5/7 designs reliably produce a protein that readily labels when BMIE is added. This work comprises the beginning steps to produce type III copper precursors with varying coordination environments and Cu-Cu distance. The design of several scaffolds with

varying Cu-Cu distances and protein environments will enable us to tease out the fundamental principles regarding dinuclear copper-oxo species.

#### 1.1.6.2 Co<sub>4</sub>O<sub>4</sub> Coordination

The OEC in photosystem II is considered one of Nature's catalytic marvels—catalyzing the oxidation of water to molecular oxygen. For decades scientists have sought to harness its catalytic potential, but due to its hydrophobic exterior, utilizing photosystem II for industrial energy production has not been feasible. A multitude of organometallic and metal-based organic nanomaterials have been proposed as light-harvesting mimics of the OEC, but to date, this remains an unsolved problem. Lack of efficient catalysis and oxidative damage are largely to blame. For example, the oxygen-evolving complex in photosystem II is 25,000 times faster and can survive a million-fold more turnovers than the coordination complex Co<sub>4</sub>O<sub>4</sub>(pyridine)<sub>4</sub>-(acetate)<sub>4</sub>, a small molecule mimic<sup>2</sup>. Additionally, it is now believed that Co<sub>4</sub>O<sub>4</sub> clusters are only capable of hydroxyl oxidation to O<sub>2</sub>. A similar complex, Co<sub>4</sub>O<sub>4</sub>(bipyridine)<sub>4</sub>-(acetate)<sub>2</sub>, has been shown not to produce oxygen at high electrochemical potentials, instead, the complex produces CO<sub>2</sub>—cannibalizing the coordinated acetates or bipyridines. Lack of oxygen production has led to the geminal-coupling mechanism hypothesis. However, QM calculations support an equally low energy barrier for a cis-coupled mechanism. The organometallic bipyridine complex, in its current form, can not survive the high potentials necessary to perform the reaction, and thus, characterization of the small molecule mimic can not be furthered without protecting the ligand environment.

For this project, we realized the potential the protein matrix might offer the  $\text{Co}_4\text{O}_4(\text{bipyridine})_4(\text{acetate})_2$  complex, protecting it from oxidative damage at higher potentials. If we can design a  $D_2$  symmetric oligomer to incorporate the tetranuclear cluster, then it would enable us to test the cis-coupling mechanism by protecting the cluster within a protein matrix. Additionally, we will exploit the symmetric scaffold to finely tune the surrounding environment to promote hydroxyl or even water oxidation.

So far, we have designed a homo-oligomeric tetramer to incorporate a non-canonical amino acid (2,2'-bipyridin-5yl)alanine (bpy) at the symmetric interface. Bpy is incorporated using the Shultz amber suppression system coupled with an orthogonal amino acid synthetase. Preliminary UV-vis and ICP-MS results have shown 1:1 (metal to monomer) binding of multiple metal ions ( $\text{Co}^{2+}$ ,  $\text{Ni}^{2+}$ ,  $\text{Cu}^{2+}$ , and  $\text{Ru}^{2+}$ ). Additionally, a silent EPR spectrum of the tetramer bound to  $\text{Co}^{2+}$  species supports a coordination structure with short metal-metal distances  $< 80\text{nm}$ , distances suggesting di or tetranuclear complex formation. The future structural characterization will determine if the various coordinated metal ions adopt a tetranuclear cubane-like complex with bridging oxo/hydroxo ligands. The complex, if realized, will be further tested for water oxidation and other redox chemistry.

### 1.1.6.3 Photo-Activatable Metalloproteins

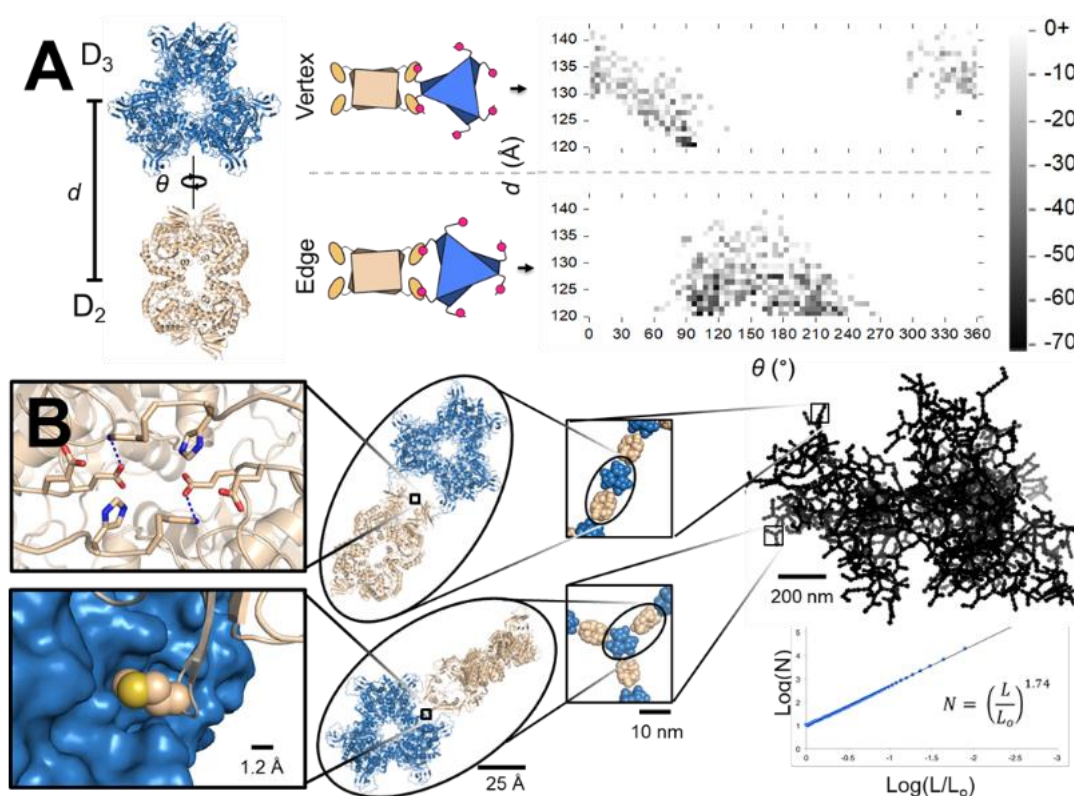
We close our discussion of SyPRIS-determined metal-binding protein interfaces with a brief look at the design of photo-activatable  $\text{C}_3$  Zn-hydrolases. This work is in collaboration with the Deiters lab who has expanded the photo-activatable azo-ncAA

library with terminally replaced imidazole/pyrazole azo moieties. For this work, we will symmetrically coordinate a tetrahedral Zn atom with the azo-pyrazole derivatives—leaving a vacant coordination site open for water coordination. Three designs were chosen and are actively being characterized by our collaborators. Additionally, we thought to repurpose the computational algorithms of SyPRIS to create an algorithm that finds the superposition of any three atoms with another three atoms. The algorithm will be used for replacing metal-coordinating histidine residues with the terminal head group of the imidazole/pyrazole-azo-benzene ncAAs to alter natural enzymes for obligate photo-activation. This work should enable the general design of metalloproteins with photo-controlled activation of enzymatic activity.

## **1.2 Introduction for Supramolecular Protein Assembly Design**

The self-assembly of proteins into organized small-scale superstructures plays an integral role in nature's biological processes<sup>14</sup>. The study of protein self-assembly has led to advances in both physiological (viral capsid and amyloid fibril) and industrial processes (liquid-liquid phase separation)<sup>15</sup>. However, many outstanding issues regarding our fundamental understanding of protein self-assembly remain. For example, one issue that remains is the complete understanding of how protein anisotropy can direct the self-assembly of proteins<sup>15</sup>. Additionally, supramolecular protein design efforts have exclusively targeted integer-dimensional organization patterns such as sheets<sup>16</sup>, lattices<sup>17</sup>, and polyhedra<sup>18</sup>. The bottom-up design of fractional-dimensional topologies had yet to be designed and studied. Fractals are ubiquitous and utilized for their high surface area to volume ratio which can lead to efficient molecular exchange (i.e., lungs and capillaries).

Furthermore, the design approaches used to obtain the supramolecular assembly have low success rates ( $\sim 10\%$ ) and rely on maintaining a specific intermolecular protein binding interface i.e., direct irreversible hydrophobic-interface design<sup>19</sup> or metal-mediated interactions<sup>19</sup>; the later relying on geometrically distinct protein shapes (cylinders, rectangular prisms, and spheres)<sup>19</sup>. A general method for the design of supramolecular protein self-assembly that does not rely on the design of novel hydrophobic interfaces and capable of simulating emergent propagation will enable designers to access fractional-dimensional topologies with higher success rates.



**Figure 2-1.**

(A) In assembly design, the energies of all favorable rotations about paired symmetry axes are modeled using (B) a stochastic propagation simulation. Atomic interactions lead to variations in attachment while remaining anisotropic to form fractal assemblies.

### 1.2.1 Statement of the Design Goal

In this half of the dissertation, we will develop a general computational design strategy that can be applied to all symmetric proteins. The strategy will include three parts: 1) develop a method for the positioning and design of flexible-fusion domains onto existing protein termini. 2) Design a computational algorithm to generate an ensemble of rigid-body placements of protein building blocks along shared symmetry axes. 3) Simulate the growing emergent patterns from a stochastic Boltzmann-weighted energy ensemble.

We will use this computational strategy to create the first fractional-dimensional supramolecular protein structures by design and show how anisotropic binding is required for fractal growth and self-similarity. Lastly, we will discuss future projects where we will combine the metal design strategy from 1.1 and the supramolecular assembly strategy from 1.2 to design 1-dimensional protein nanowires capable of long-range electron transfer.

### 1.2.2 General Design Strategy for Supramolecular Protein Assembly

When developing a general design strategy, we wanted to highlight the inherent benefits of design with symmetry. Using symmetry for supramolecular assembly has well-documented successes<sup>18</sup>, including the design of capsid-like (wire-frame) closed-form protein assemblies<sup>19</sup>. Despite these successes, the strategies that rely on symmetry have only targeted the design of integer-dimensional protein assembly (rods, sheets, crystals). Symmetric propagation greatly reduces computational design, even to a single interface. However, the design of integer-dimensional assemblies requires exacting anisotropic

interface interactions—any less than the perfect angle or distance would result in unrealized assembly or, worse, protein agglomerates. In order to make assembly by design generalizable, self-assembly should be achievable reliably with any two-component protein system of interest without the need for explicit interface design. One strategy, ubiquitous in Nature, is to incorporate high-affinity water-soluble fusion domains (protein-peptide or peptide-peptide) interactions that fold reliably independent of their fused protein counterparts. High-affinity fusion domains will replace the need to design novel interfaces that typically have a 10% success rate. In this dissertation we will discuss the use of the stimulus-responsive SH2 domain and its target peptide partner. Our method of using utilizing high-affinity fusion domains attached to symmetric proteins should enable the general design of future protein assemblies. The following methods are covered extensively in Chapter 8.

#### 1.2.2.1 Symmetric Intermolecular Rigid-Body Ensemble Generation

Modifications to symmetric proteins are copied to all protein subunits. For example, fusing a domain onto the exposed amino-terminus of a tetramer will enable that tetramer to have four copies of the fusion domain. Proteins often exhibit two major symmetries: cyclic (C) and dihedral (D). Cyclic proteins have just one axis, and units are copied about that axis at  $360/n$  intervals where  $n$  is the number of subunits in the oligomer ( $C_3$  proteins have three subunits at 120-degree intervals about the  $C_3$  axis). More complicated are the  $D_n$  symmetric proteins—which have 1- $C_n$  axis and  $n$ - $C_2$  axes (For example, a  $D_4$  protein has 1  $C_4$  axis and 4- $C_2$  axes). Another way to imagine D-symmetric proteins is to create a C-symmetric "sandwich", the major axis is maintained, but you have just created multiple symmetric dimers about the central axis at  $360/n$  intervals. Propagation along the major

axis will typically lead to wire formation, while propagation in the orthogonal plane of the major axis will lead to sheets and 2-dimensional planes. 3-D structures such as crystals can be attained via propagation along the central axis and the orthogonal plane. Emergent properties are harder to predict, but 3-D structures may also be achieved by introducing rotations out of the orthogonal plane for subsequent assembly units in a two-body system. Our approach is to create an ensemble of intermolecular two-body interfaces by varying the rigid body position of one protein oligomer with another. For this dissertation, we will discuss the ensemble of positions generated for  $D_3$  and  $D_2$  symmetric proteins, which, by definition, share  $C_2$  axes in common. The design approach aligns proteins along their shared  $C_2$  symmetry axis and creates a library of feasible (non-clashing) rigid body positions (Fig. 2A). The library is achieved via rotations and translations of one component about or along paired symmetry axes.

#### 1.2.2.2 RosettaMatch for Domain-Fusion *in silico*.

With an ensemble of novel intermolecular protein interfaces, we want to locate ensemble members that present their sequence termini such that the simultaneous fusion of the domain-peptide partners can be realized on each protein subunit while preserving their high-affinity contact. To this end, our approach utilizes a computational tool in Rosetta called RosettaMatch (purposed for finding small molecule binding sites in proteins) for locating a geometrically acceptable placement of the high-affinity fusion partners between the protein components in the two-component protein library. The placement of the fusion partners is guided by geometric backbone geometries derived from the non-redundant protein library. This matching algorithm is applied to all inter-component ensemble members until a suitable rigid-body placement is found. The approach was first

utilized to model missing crystallographic density and position a macrocyclic inhibitor (Chapters 6 and 7).

#### 1.2.2.3 Rosetta FastDesign of Novel Interfaces

When a suitable match for the domain-peptide fusion is found, we covalently link the protein domains and perform RosettaDesign on all novel interfaces. All but the linker backbone atoms remain fixed during design. During design, residues that would otherwise clash in our hypothetical structure are removed, and new amino acid substitutions that favor that rigid-body placement are introduced—such as hydrogen bonding or hydrophobic interactions (Fig. 2B). Spurious space-filling substitutions, an artifact of Rosetta's implicit water model, are reverted to their native identities.

#### 1.2.2.4 Stochastic Propagation Simulation

Propagation of any protein assembly may result in emergent topological properties. Unlike reported examples where the design is driven by topology, here, the fusion domains and the inter-component placement drive propagation and impose geometric constraints on the assembly. Determining emergent topological patterns is an essential step to determine the size, shape, porosity, and density of our designed symmetric assemblies. For this, we developed a coarse-grained stochastic simulation that utilizes Boltzmann-weighted probabilities of the ensemble members (Figure 2A right) to propagate a structure with some "sticking probability" (Figure 2B right).

### 1.2.3 Fractional-Dimensional Protein Self-Assembly by Design

The approach outlined in the previous section was used to generate the first fractional-dimensional supramolecular protein assembly by design. We wanted to show that the

approach could be broadly applicable to any symmetric protein partners and so we chose to use two enzymes in the atrazine degradation pathway (AtzA and AtzC), D<sub>3</sub>, and D<sub>2</sub> symmetric proteins. A reversible SH2 domain-peptide pair was chosen as the paired fusion molecules to be inserted. Specific rotations about the paired symmetry axes (C<sub>2</sub>) lead to integer-dimensional self-assembly. For example, a plane could be produced if the rotation about the C<sub>2</sub> axis was forced to remain with a single angle in the set {0, 90, 180, 270}°. A 3-dimensional lattice could be designed in the same way by forcing the design to adopt a rotation in the set  $180 \pm \{35.25, 54.75, 125.25, 144.75\}^\circ$ . Instead, we chose to accept amino-substitutions that allowed fusion placements to adopt favorable energy across *multiple* rotations about the paired C<sub>2</sub> symmetry axis. In this way, we anticipated that although the binding of the two components would be anisotropic (divalent connection across the C<sub>2</sub>-symmetry axis), the emergent topology of the supramolecular assembly would be driven by statistically stochastic thermodynamic interactions. To model the possible assemblies, we created a coarse-grained stochastic assembly growth simulation program that patterns layers of symmetric components using a Boltzmann-weighted energy distribution to calculate rotations about the C<sub>2</sub> symmetry axes for subsequent components. The simulation was used to observe the emergent topologies generated from atomic interactions. Not only did the simulations and the experimentally characterized (cryoEM) protein assembly form with a fractional dimensional topology self-similar across multiple length scales, the comparative analysis showed that the simulation under certain parameters i.e., high temperature (assembly at RT) and high sticking probability (4nM binding of SH2-peptide partner) accurately predicted the fractal dimension, intercomponent distance distribution, and the average distribution of

connected components. To prove that the fractal protein assembly was indeed a product of stochastically varied anisotropic binding events, we added a 10x-(GSS) long flexible linker at the fusion site and observed globular or amorphous assembly formation expected for isotropic binding. This study provided clear evidence that self-similar fractal topologies could be achieved with stochastic anisotropic binding, without the need to design extensive and limiting hydrophobic interfaces. Furthermore, the approach can be utilized for any symmetric protein components of interest. The complete characterization of the fractal assembly showed increased resistance to physical stressors (shaking speed and high temperature) but was confirmed to be a product of all assembly formation. However, the fractal was shown to be a better topology than the globular assembly for the sequestration of IgG antibodies and other relatively large biomolecules. These results reinforce that new functions can arise from self-assembly of proteins and may be applied in bioremediation, drug-delivery, and other industrial processes. This work is covered in Chapter 8.

#### 1.2.4 The Future of Protein Self-Assembly and MNMP Design

Future efforts utilizing the methods outlined in the dissertation have already begun. One example of this is the design of a protein nanowire capable of long-range electron transfer. To achieve this, a small D<sub>2</sub> homo-oligomeric 4-helix bundle was selected as the core repeating unit of the nanowire. The wire assembly was computationally designed using a 2-helix hetero-dimer fusion. Simultaneously, the geometric center of the bundle was determined computationally by SyPRIS as a geometrically ideal location for a Fe<sub>4</sub>S<sub>4</sub> cluster coordinated by four Cys residues (a single propagated residue substitution). The

computationally derived intercomponent distance of each cluster based on simulated propagation is expected to be between 15-20Å (at the edge of Fe<sub>4</sub>S<sub>4</sub> cluster electron transfer capabilities). If the electron transfer wire is realized, symmetry will be further exploited to begin understanding the impact that first and second shell interactions have on long-range electron transfer. This project will not be covered in this dissertation but is meant to inform how the methods developed here are supporting future research endeavors.

### 1.3 References

- 1 Pal, C., Balazs, P., Lercher, & M.J. An integrated view of protein evolution. *Nature Review Genetics* 7:337-348 (2006).
- 2 McCool N.S, Robinson, D.M., Sheats, J.E., & Dismukes, C. A Co<sub>4</sub>O<sub>4</sub> "Cubane" Water Oxidation Catalyst Inspired by Photosynthesis. *JACS* **133**(30):11446-11449 (2011).
- 3 Leaver-Fay, *et al.* A. ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**:545-574 (2011).
- 4 Guffy, S.L., Der, B.S., & Kuhlman, B. Probing the minimal determinants of zinc-binding with computational protein design. *PEDS* **29**(8): 327-338 (2016).
- 5 Dudev, T., & Lim, C. Metal Binding Affinity and Selectivity in Metalloproteins: Insights from Computational Studies. *Annu. Rec. Biophys.* **37**:97-116 (2008).
- 6 Mills, J.H., *et al.* Computational design of an unnatural amino acid-dependent metalloprotein with atomic level accuracy. *JACS* **135**(36):13393-13399 (2013).

- 7 Mills, J.H., *et al.* Computational design of a homotrimeric metalloprotein with trisbipyridyl core. *PNAS* **113**(52):15012-15017 (2016).
- 8 Davies, R.R., & Distefano, M.D. A Semisynthetic Metalloenzyme Based on a Protein Cavity That Catalyzes the Enantioselective Hydrolysis of Ester and Amide Substrates. *JACS* **119**(48):11643-11652 (1997).
- 9 Laganowsky, A., *et al.* An approach to crystallizing proteins by metal-mediated synthetic symmetrization. *Prot. Sci.* **20**(11):1876-1890 (2011).
- 10 Choma, C.T., *et al.* Design of a heme-binding four-helix bundle. *JACS* **116**(3):856-865 (1994).
- 11 Richter, F., Leaver-Fay, A. Khare, S.D., Bjelic, S. & Baker, D. *De novo* enzyme design using Rosetta3. *PloS ONE* **6**(5): e19230 (2011).
- 12 Hansen, W.A., & Khare, S.D. Computational Design of Multinuclear Metalloproteins Using Unnatural Amino Acids. *Springer Protocols: Methods in Computational Protein Design* **13**:1414 (2016)
- 13 Hansen, W.A., Khare, S.D. Benchmarking a computational design method for the incorporation of metal ion-binding sites at symmetric protein interfaces. *Protein Science* **26**(8):1584-1594 (2017).
- 14 Ahnert, S.E., Marsh, J.A., Hernandez, H., Robinson, C.V., & Teichmann, S.A. Principle of assembly reveal a periodic table of protein complexes. *SCIENCE* **350**:6266 (2015).

- 15 McManus, J.J., Charbonneau, P. Zaccarelli, E. & Asherie, N. The physics of protein self-assembly. *Curr Opin Colloid In* **22**:73-79 (2016).
- 16 Suzuki, Y., *et al.* Self-assembly of coherently dynamic, auxetic, two-dimensional protein crystals. *Nature* **533**:369-373 (2016).
- 17 Sinclair, J.C., Davies, K.M., Venien-Bryan, C. & Noble, M.E.M. Generation of protein lattices by fusing proteins with matching rotational symmetry. *Nat Nanotechnol* **6**:558-562 (2011).
- 18 King, N.P., *et al.* Computational Design of Self-Assembling Protein Nanomaterials with Atomic Accuracy. *SCIENCE* **336**(6085):1171-1174 (2012).
- 19 Salgado, E.N., Faraone-Mennella, J., & Tezcan, F.A. Controlling protein-protein interactions through metal-coordination: Assembly of a 16-helic bundle protein. *JACS* **129** (44):13374-13375 (2007).

## **Chapter 2: Computational Design of Multinuclear**

### **Metalloproteins Using Unnatural Amino Acids**

#### **2.1 Preface**

A version of this chapter has been published in Springer Protocols: Computational Design of Ligand Binding Proteins and is formatted in the journal style.

#### **2.2 Summary**

Multinuclear metal ion clusters, coordinated by proteins, catalyze various critical biological redox reactions, including water oxidation in photosynthesis, and nitrogen fixation. Designed metalloproteins featuring synthetic metal clusters would aid in the design of bio-inspired catalysts for various applications in synthetic biology. The design of metal ion-binding sites in a protein chain requires geometrically constrained and accurate placement of several (between three and six) polar and/or charged amino acid sidechains for every metal ion, making the design problem very challenging to address. Here, we describe a general computational method to redesign oligomeric interfaces of symmetric proteins to create novel multinuclear metalloproteins with tunable geometries, electrochemical environments, and metal cofactor stability *via* first and second shell interactions.

The method requires a target symmetric organometallic cofactor whose coordinating ligands resemble the sidechains of a natural or unnatural amino acid, and a library of oligomeric protein structures featuring the same symmetry as the target cofactor. Geometric interface matches between target cofactor and scaffold are determined using a

program that we call Symmetric Protein Recursive Ion-cofactor Sampler (SyPRIS). First, the amino acid bound organometallic cofactor model is built and symmetrically aligned to the axes of symmetry of each scaffold. Depending on the symmetry, rigid body and inverse rotameric degrees of freedom of the cofactor model are then simultaneously sampled to locate scaffold backbone constellations that are geometrically poised to incorporate the cofactor. Optionally, backbone remodeling of loops can be performed if no perfect matches are identified. Finally, the identities of spatially proximal neighbor residues of the cofactor are optimized using RosettaDesign. Selected designs can then be produced in the laboratory using genetically incorporated unnatural amino acid technology, and tested experimentally for structure and catalytic activity.

## 2.3 Introduction

Much progress has been made in the last two decades toward the *de novo* design of novel metalloproteins[1–9], where the guiding principle is the simultaneous placement of two or more metal coordinating sidechain groups from naturally occurring amino acid residues: Cys, Asp, Glu, and His. However, successful design attempts have been dominated mainly by *mononuclear* (a single metal ion per designed protein) insertions into a single type of scaffold – the geometrically well-defined alpha helical bundles[3]. One of the challenges while designing a *multinuclear* (metal ion site composed of two or more metal ions) metalloproteins is the need to incorporate multiple sidechain coordinating groups in close spatial proximity in a single protein — placing exacting constraints on design. Another challenge is the design of the electrostatic environment of

the metal ions, which has a significant impact on the stability of the highly charged cofactor and the associated catalytic activity.

Computational algorithms could, in principle, aid in addressing both challenges. We previously developed an algorithm that utilized the metal-chelating unnatural amino acid 2,2'-[bipyridyl]alanine (BPY)[10, 11] for designing mononuclear metal binding sites [9]. The algorithm uses RosettaMatch [12] to search combinatorially, in a given protein scaffold (typically a single chain), for a constellation of backbone structures that can support the multiple (~3-6) sidechain metal-chelating functional groups in the appropriate coordination geometry. The use of BPY simplified the combinatorial design problem as, unlike any natural amino acid sidechain, the bipyridyl moiety contributes two metal ligands from the same amino acid sidechain. Metalloproteins featuring BPY with His and Asp/Glu residues were designed and their crystallographic structure demonstrated close agreement with the design model. However, this algorithm is limited by its combinatorial complexity and is not applicable, practically, to construct multinuclear metal-binding sites.

Here, we describe an approach to computationally design incorporation of a symmetric multinuclear metallo-cofactor via integration into a similarly symmetric protein scaffold (Fig. 1). For this task, we have developed a matching algorithm, Symmetric Protein Recursive Ion-cofactor Sampler (SyPRIS), and implemented it in Python. This algorithm allows expanding metalloprotein design to scaffolds other than alpha helical bundles, as well as gaining access to a greater variety of symmetric multinuclear cofactors such as iron-sulfur clusters, and cubane complexes. We illustrate the method by describing the incorporation of the  $D_2$  symmetric cobalt-oxygen cube-like cofactor (Co-Cubane) [13–

20]. This cofactor is a mimic of the water oxidation center in Photosystem II, and features four bipyridyl moieties co-ordinating four Co-ions, respectively. Though Co-cubane is used as an example, the method is generally applicable to incorporate all types of cofactors of either C or D symmetry within any complementary symmetric scaffold. Theozyme [21] matches generated from SyPRIS can be further designed with the enzyme design modules in the Rosetta macromolecular modeling software [12, 22–25].

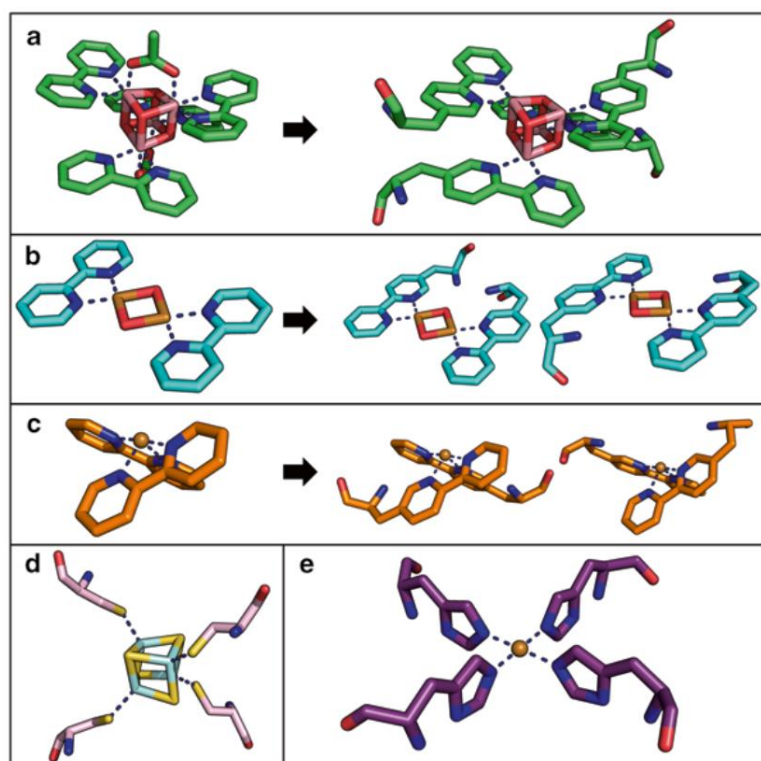


Figure 3-2. Figure 1. Several target cofactors that this method was intended to implement using scaffolds of various symmetries.

(a) Co<sub>4</sub>O<sub>4</sub>(Ac)<sub>2</sub>(bipyridine)<sub>4</sub> converted from CCDC crystal structure to noncanonical amino acid-bound model featuring D<sub>2</sub> symmetry. (b) Cu<sub>2</sub>(OH)<sub>2</sub>(bipyridine)<sub>2</sub> converted to models featuring C<sub>2</sub> symmetry. (c) CuOH(bipyridine)<sub>2</sub> converted to models featuring C<sub>2</sub> symmetry. (d) Fe<sub>4</sub>S<sub>4</sub>(Cys)<sub>4</sub> cluster featuring D<sub>2</sub> symmetry. (e) Cu(OH)<sub>2</sub>(His)<sub>4</sub> featuring C<sub>4</sub> symmetry.

## 2.4 Methods

2.4.1 The general pipeline for the method (Fig. 2A) includes the following steps:

1. Generate and standardize a symmetric scaffold library (Fig. 3B).
2. Prepare a target cofactor for symmetric insertion (Fig. 3C).
3. Use SyPRIS to identify inverse rotamer positions suitable for design (Fig. 3D).
4. Perform kinematic loop closure on residue matches that reside within a loop secondary loop structure (Figs. 3E and 3F).
5. Design the oligomeric interface with constraints (Fig. 3G).
6. Revert extraneous residue mutations to favor wildtype sequence.
7. Experimental validation through protein expression, purification, and crystallization [are not discussed here].

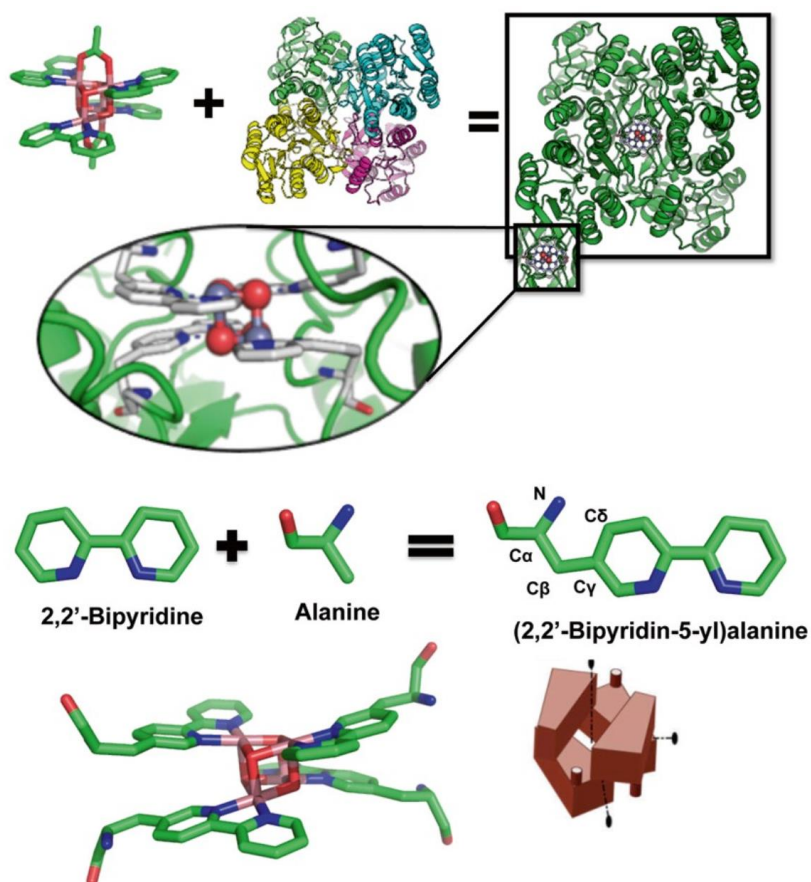


Figure 4-2. Figure 2. Method overview, incorporation of a  $\text{Co}_4\text{O}_4(\text{Ac})_2(\text{bipyridine})_4$  cofactor with noncanonical amino acids into a  $D_2$  symmetric scaffold.

#### 2.4.2 Generate and standardize symmetric scaffold library

Potential protein scaffold candidates are selected from the RCSB protein databank to feature a given symmetry in the oligomeric protein i.e.,  $D_2$ ,  $C_{2,3,4,\dots}$ . Search parameters include symmetry type, chain stoichiometry, expressibility in *E. coli*, 90% sequence identity threshold,  $<3.0\text{\AA}$  resolution (for structures determined by X-ray crystallography). From these constraints, a raw scaffold library is generated. More than 70% of the scaffold files generated in this way contain asymmetries in the form of incomplete chains—due to missing electron density in the crystal structures. In order to use the symmetry package of the Rosetta suite, all input files must be composed of chains that are equal in both residue length and residue type. To correct the intrinsic asymmetries, a hybrid Smith-Waterman local alignment is performed on all combinations of chains, removing residues absent from other chains, until a single converging monomeric sequence and all its symmetric partner protomers in the structures are found.

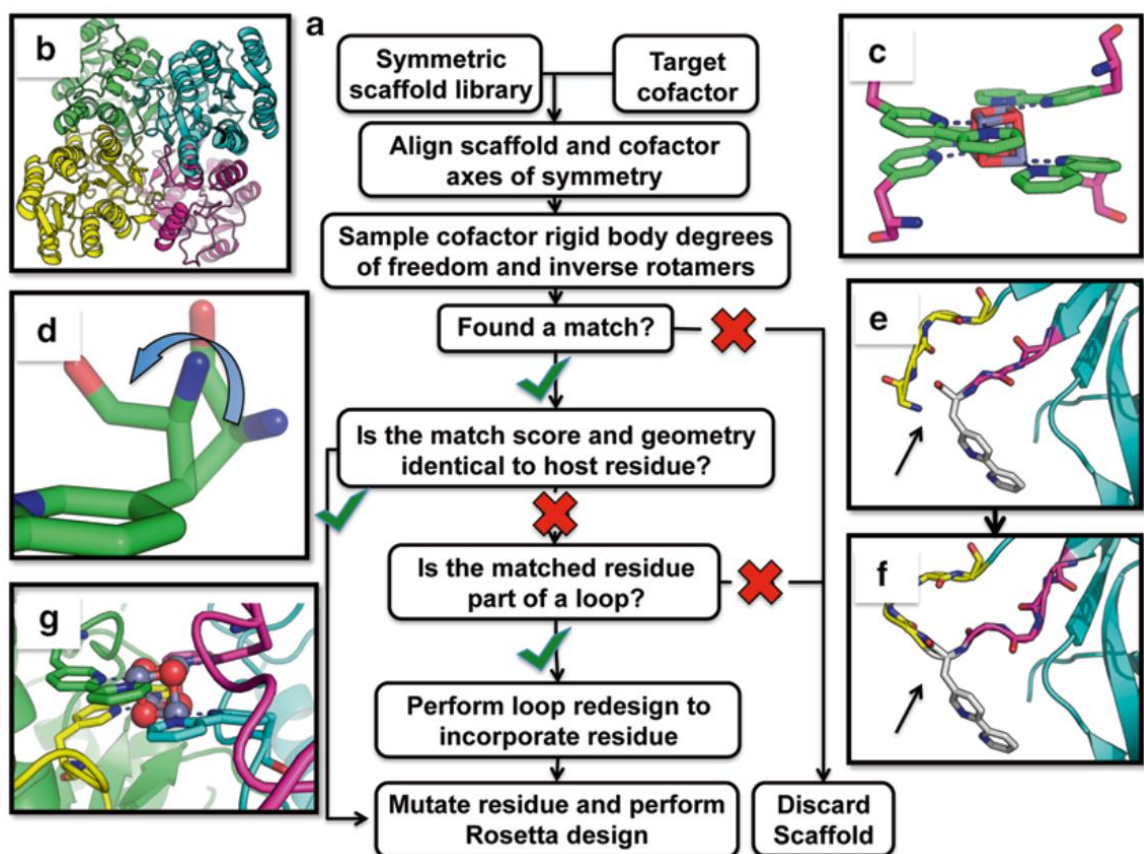


Figure 5-2. Figure 3. SyPRIS design flowchart.

(a) SyPRIS flow chart starting from generating a scaffold library and ultimately ending in a designable or discarded match. (b) An example scaffold, part of a library, will be considered by SyPRIS for the incorporation of a target cofactor. (c) A target cofactor, in this case, an oxocobalt cubane coordinated by bipyridine ligands, has been modified with the appended magenta atoms creating a noncanonical amino acid. (d) The rotameric degrees of freedom for the atoms comprising the new backbone are sampled recursively with a chi distribution file (or exhaustively if desired) and compared to that of nearby backbone residues of the scaffold. (e) If the matched residue is part of a loop and the match was not geometrically identical, the loop is remodeled. (f) Three residues upstream and downstream of the translated backbone position are remodeled using Generalized KIC in Rosetta. (g) A fully designed oligomeric interface showing incorporated cofactor.

### 2.4.3 Target cofactor

Cofactors of interest include organometallic compounds containing ligands that resemble either canonical amino acids or previously characterized noncanonical amino acids. PDB

files are generated for cofactors of interest using their crystal structures and, where needed, the programs Mercury 3.5 and ConQuest 1.17 from the Cambridge Crystallographic Database (CCDC). Small structural changes may be applied to the supplied atom positions to reduce asymmetries within the X-ray crystallographic models. If necessary, backbone atoms are appended to each symmetric ligand and all dihedrals are set to a default 0.0 degrees prior to matching. To identify dihedral positions acceptable for each cofactor, an ensemble is generated of all dihedral rotations while simultaneously performing internal atomic clash checks. Dihedral rotations that pass the clash check are stored and plotted against each subsequent dihedral rotation within a heatmap. Preferred geometries are classified as regions of the heatmap with the highest bin density at a determined threshold. These geometric constraints are then converted into a "chi distribution" file necessary for the Symmetric Protein Recursive Ion Sampler (SyPRIS). A chi distribution file depicts the four atoms participating in a dihedral rotation, a range of values between which to sample, and the degree with which to iterate. A Rosetta parameter file, that stores information about the asymmetric unit of the multinuclear cluster (i.e. one Co ion and one oxygen atom for the Co-cubane, one Fe and one S atom for an iron-sulfur cluster), is defined for integration within the Rosetta suite during design. Lastly, a Rosetta enzyme design constraints file, that adds an energy term favoring the coordination geometry between ligand and complex, is generated to more accurately determine the energy of the integrated cofactor.

#### 2.4.4 Symmetric Protein Recursive Ion Sampler (SyPRIS)

With the scaffold set and cofactor model in place, the following steps are utilized in finding symmetric matches between the cofactor coordinated to a UAA and the protein scaffold:

#### 2.4.4.1 Align scaffold and cofactor axes of symmetry

1. The axis of symmetry for the scaffold protein and each cofactor are determined by finding the Eigenvector and Eigenvalues—multiplying the coordinate matrix by its transpose matrix. Consequently, this creates unit vectors for each set of coordinates and supplies the principle rotational axes defined as the Eigen minimum, maximum and their orthogonal cross product. In C-symmetry proteins the Eigen minimum and maximum can each be the target axis of symmetry. To correctly identify the axis of symmetry in a C-system, the midpoint of all symmetric Ca atoms is generated and the average of all vectors connecting atoms to the origin becomes the symmetric axis.
2. Translate all Cartesian atoms of all files so that the axis of symmetry origin from the scaffold and each model lie on a theoretical (0,0,0) origin.
3. Align the axes of symmetry of the complex so that the Eigen maximum and Eigen minimum are aligned with that of the given scaffold (Fig. 4B). In C-symmetry, the Eigen minimum of the cofactor is aligned to the midpoint average vector generated in step 1.

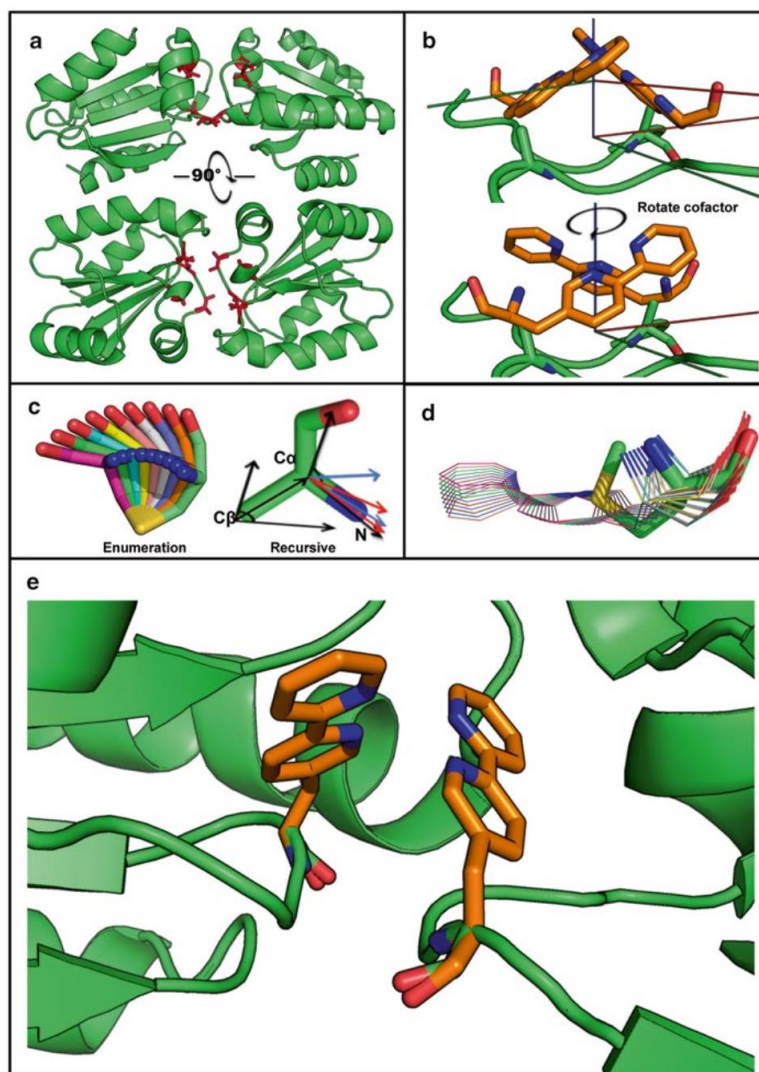


Figure 6-2. Figure 4. Alignment of backbone atoms.

(a) Residues that satisfy user-specified distance from the symmetric axis highlighted in red sticks. (b) Rigid body rotation about the symmetric axis to align symmetric axes. (c) Pictorial view of the enumerative exhaustive backbone sampling (left). Schematic view of the recursive atom placing algorithm for direct matching (right). (d) An ensemble of backbone positions generated via the recursive method. (e) A matched cofactor output from SyPRIS ready for Rosetta Design.

4. If the input features C-symmetry, SyPRIS will locate the midpoint of the C $\beta$  atoms of the cofactor and translate to the midpoint of each protein C $\beta$  combination that is within  $\pm <\text{user input (default = 1.0)}> \text{\AA}$  of the cofactor C $\beta$  radii (Fig 3A). The

cofactor is then rotated about the plane of symmetry until the C $\beta$  atoms of both the cofactor and protein are aligned (Fig 3B). Each rotational/translational position unique to a residue subset will store the lowest atom magnitude difference position as well as two other rotational positions clockwise and counterclockwise to the aligned atoms within a <user input (default = 1.0)>Å direct distance. The four unaligned positions will be stored to generate an ensemble of positions and dihedrals further starting from step 6, below.

5. If the input features D-symmetry, SyPRIS will perform 90° and 180° rotations of the cofactor about the vectors that correspond to each of the defined symmetric axes. Each rotational position will be further sampled in step 6.

#### 2.4.5 Sample inverse rotamers

1. A cofactor to scaffold backbone clash check is performed by determining distances between all heavy atoms of the cofactor not included in the chi distribution file and the backbone heavy atoms of nearby residues (not including the residue making the match  $\pm$  one residue position proximal in sequence). Any distances to heavy atoms less than <user input (default = 2.8 Å)> are considered clashes, and discarded.
2. For each unique cofactor rotation, cofactor backbone atoms (branches) are rotated within the range of values about the bonds defined by the atoms in the chi distribution file.
3. To score a given rotation, a vector is produced from the last stationary atom (LASA) to the first atom changing location (FACL). For example, while rotating about a chi1 bond of BPY UAA the LASA is the alpha carbon and the FACL would be the

backbone nitrogen atom. The vector produced by the LASA and FACL of the cofactor is compared to that of the scaffold. The angle difference is calculated as an AngleLog:

$$\text{AngleLog} = \log(\Sigma \Delta [ (\cos^{-1} ( \langle \text{xyz} \rangle \bullet \langle \text{xyz}' \rangle / \| \text{xyz} \| * \| \text{xyz}' \| ) n / 20 * n ) ] )$$

Where n is the number of compared vectors and a value of zero is an average deviation of 20° across all n vectors. To further score a matched position, the magnitude of the cofactor FACL to the compared scaffold atom is calculated. The default threshold for AngleLog and atom-magnitude is <user input (default = 0.0)> and <user input (default = 0.8)>Å respectively.

4. Enumerative Sampling. A predefined ensemble of inverse rotameric states is stored within one cofactor file. Each state is sampled exhaustively (Fig. 4C left).
5. Recursive sampling. For any range of values tested in the chi distribution file, the best scoring rotation (as long as it meets the thresholds) is stored along with the best adjacent rotation. Recursive ½ angles are sampled within this range to minimize to the best solution. The algorithm to locate new half dihedrals:

$$\text{A) } (\varphi_o + \varphi_n / 2)^n \quad \text{or} \quad \text{B) } (\varphi_{n-1} + \varphi_n / 2)^n$$

Where n is the number of half angles calculated as set by the user,  $\varphi_o$  is the first dihedral (best scored), and  $n=1$  is the best scoring adjacent dihedral. SyPRIS starts with the algorithm in A. If two of the newly calculated half angles score better than the original dihedral, the B algorithm takes over for subsequent tests. Only the  $\varphi_o$ ,

$\phi_1$ , and  $\phi_n$  ( $n=\max$ ) FACL rotated branches will be stored to sample a wider ensemble of positions (Fig. 4C right). This algorithm occurs for each subsequent torsion angle at all stored positions ( $3^{\#}$  of chis). Therefore a cofactor with three chis featuring  $D_2$  symmetry will store 27 positions (with tunable tolerance) at a given rotation. A  $C_2$  cofactor with the same number of chis will store up to five times this many positions due to the rigid body rotational degrees of freedom (Fig. 4D).

6. For both the Recursive and Enumerative methods, final matches are determined by the scoring the average AngleLog and RMSD over all FACL atom positions as defined in step 8 (Fig. 4E).
7. A table for each protein is generated containing all the intrinsic properties of the ion cluster at a given match—model number and rotation about an axis. The table also includes the residue matched within the scaffold, the average Anglelog score, each individual Anglelog for all chains, the RMSD for all compared atoms, and the scaffold name. If an exact match is found (priority 1 designs), the scaffold will be mutated at the given residue position and passed to Rosetta design. All other matches are subjects for the KIC procedure (priority 2 designs).

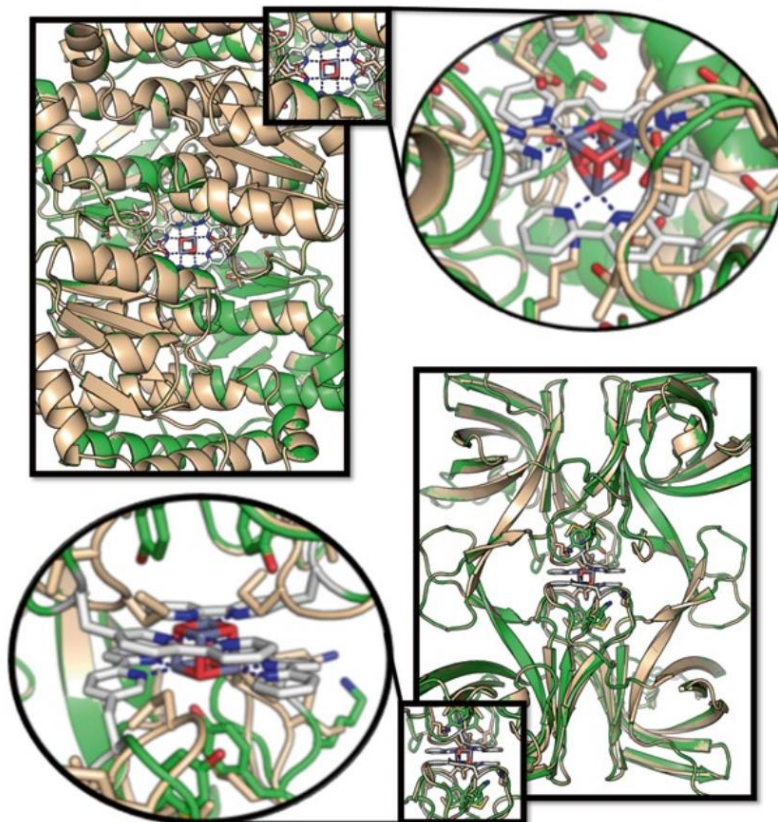


Figure 7-2. Figure 5. SyPRIS design output.

Two designs involving a catalytic D2 symmetric organometallic cofactor ( $\text{Co}_4\text{O}_4(\text{Ac})_2(\text{bipyridine})_4$ ). The noncanonical amino acid bipyridine is incorporated on one chain, forming the cofactor upon oligomerization. The design protein (green and white) is compared to the wildtype scaffold (wheat). Sticks represent mutations positions.

#### 2.4.6 Kinematic Loop Closure (KIC)

This pre-design method takes the tables generated by SyPRIS and locates the preferred residues for replacement with the ligand-like amino acid within the protein scaffold. The secondary structure of that residue with  $\pm$  <user input (default = 3)> residues is determined based on Ramachandran preferred angles of phi and psi using a standard DSSP check. If the query within the scaffold is a loop region the scaffold is accepted as designable; otherwise, if the region is helical or forms beta sheets the scaffold is rejected. The scaffolds containing loops at match locations are then subjects of programs that:

1. Take the scaffold and corresponding model as arguments.
2. Translate the backbone coordinates of the matched residue on the scaffold to the location of the model to ensure the exact match (generally changing atom positions by 0.5Å across the entire residue).
3. Generate a coordinate constraint file (See note 2) of the heavy atoms comprising the multinuclear cluster in the model corresponding to chain A for use during design. A coordinate constraint (CST) file contains coordinates that ensure that the metal cluster atoms do not change positions during design.
4. Generate two 'loops' files (upstream and downstream of the matched residue) specific to each scaffold and matching residues necessary for performing KIC. The loops file contains information for which residues backbones will be sampled to make connection to another endpoint residue (i.e. remodeling the upstream or downstream loop about the ligand-like residue).
5. Utilizing a Rosetta Generalized KIC [26, 27] , the four residues upstream and downstream are remodeled to accommodate the new position of the matched residue (step II). The remodeling includes sampling of backbone phi and psi angles while progressively closing the chain break. More details can be found in Kortemme et al.
6. A deterministic de novo loop is generated for each use of Generalized KIC.
7. Generated loops are evaluated based on void formation, electrostatic repulsion, etc.

#### 2.4.7 Rosetta Design

All redesigned loop scaffolds that pass are subject to four rounds of rotamer sampling followed by gradient-based minimization of the side chain and backbone atoms. Design

and repack shells are defined as residues with C $\alpha$  atoms within 12Å and 16Å radii, respectively, about the matched residue. The design shell specifies that all residues within the shell excluding the metal cofactor and UAA will be allowed to mutate to other more favorably scoring residues. Residues within the repack shell sample their rotameric preferred side chain conformations while keeping their identity fixed. The talaris2013 symmetric score function with constraints is used to evaluate the states of the protein during design. The coordinate constraint file generated in Step 3 of KIC is used to force the ligand-like residue into a conformation conducive for coordinating the ions of the cofactor. The symmetry definition file generated in stage 2 was used to copy any change made on the master unit to all slave units, as defined by Rosetta symmetry. Backbone minimization is allowed for residues that are part of the UAA-containing loop and nearby residues. Heavy coordinate constraints are placed on the scaffold to only allow movement of backbone atoms if necessary due to redesigned loop clashes. Final designs are chosen by low backbone RMSD of the design shell, smallest change to void volume, and favorable energies of the interaction of the design shell residues with the cofactor (See notes 3 and 4). Lastly, reversion is made on extraneous residues (See note 5) to favor the wild type sequence, and the protein is ready for expression (Fig. 5).

#### 2.4.8 Notes

1. The Rosetta force field, as other molecular mechanics force fields, does not accurately model interactions of protein functional groups with metal ions. Therefore, it is necessary to treat these interactions with restraints. The weights used in the restraints will be system dependent, but in the final models, one should end up with a metal site geometry similar to the one from the starting crystal structure with some small deviation.

If the metal site is completely distorted, the weights of the restraints should be increased to keep the geometry fixed.

2. Another metric that is currently evaluated by human intuition in our protocol is that access to small ions/substrates to the metal site has not been blocked by new mutations introduced in the design protocol. Conformational changes upon substrate binding are not modeled and system-dependent knowledge of the dynamics of the closure and opening of the active site should be kept in mind when either picking out scaffolds for design and evaluating designs by inspection.

3. Many substitutions can be introduced but as a designer one should also make sure that the initial protein scaffold can accommodate these changes in the absence of any substrate; otherwise the enzyme will either not express or be unfolded. In particular, we paid special attention to the maintenance of the symmetric interface of the oligomer in question.

4. Chemical intuition is almost always required to evaluate the goodness of designs.

## 2.5 References

1. Ghosh D, Pecoraro VL (2004) Understanding metalloprotein folding using a de novo design strategy. *Inorg Chem* 43:7902–7915. doi: 10.1021/ic048939z
2. Hellinga HW (1996) Metalloprotein design. *Curr Opin Biotechnol* 7:437–441. doi: 10.1016/S0958-1669(96)80121-2
3. Peacock AFA (2013) Incorporating metals into de novo proteins. *Curr Opin Chem Biol* 17:934–939. doi: 10.1016/j.cbpa.2013.10.015

4. Zastrow ML, Pecoraro VL (2013) Designing functional metalloproteins: From structural to catalytic metal sites. *Coord Chem Rev* 257:2565–2588. doi: 10.1016/j.ccr.2013.02.007
5. Lu Y, Yeung N, Sieracki N, Marshall NM (2009) Design of functional metalloproteins. *Nature* 460:855–862. doi: 10.1038/nature08304
6. Grzyb J, Xu F, Weiner L, et al. (2010) De novo design of a non-natural fold for an iron-sulfur protein: Alpha-helical coiled-coil with a four-iron four-sulfur cluster binding site in its central core. *Biochim Biophys Acta - Bioenerg* 1797:406–413. doi: 10.1016/j.bbabi.2009.12.012
7. DeGrado WF, Summa CM, Pavone V, et al. (1999) De novo design and structural characterization of proteins and metalloproteins. *Annu Rev Biochem* 68:779–819. doi: 10.1146/annurev.biochem.68.1.779
8. Degrado WF, Summa CM, Pavone V, et al. (1999) De Novo Design and Structural Characterization of Proteins. *Biochemistry* 779–819.
9. Mills JH, Khare SD, Bolduc JM, et al. (2013) Computational design of an unnatural amino acid dependent metalloprotein with atomic level accuracy. *J Am Chem Soc* 135:13393–13399. doi: 10.1021/ja403503m
10. Liu CC, Schultz PG (2010) Adding new chemistries to the genetic code. *Annu Rev Biochem* 79:413–444. doi: 10.1146/annurev.biochem.052308.105824
11. Imperiali B, Fisher SL (1991) (S)-u-Amino-2,2'-bipyridine-6-propanoic Acid: A Versatile Amino Acid for de Novo Metalloprotein Design. *J Am Chem Soc* 113:8527–8528. doi: 10.1021/ja00022a053
12. Richter F, Leaver-Fay A, Khare SD, et al. (2011) De novo enzyme design using Rosetta3. *PLoS One* 6:1–12. doi: 10.1371/journal.pone.0019230
13. Smith PF, Kaplan C, Sheats JE, et al. (2014) What determines catalyst functionality in molecular water oxidation? dependence on ligands and metal nuclearity in cobalt clusters. *Inorg Chem* 53:2113–2121. doi: 10.1021/ic402720p
14. Li X, Clatworthy EB, Masters AF, Maschmeyer T (2015) Molecular Cobalt Clusters as Precursors of Distinct Active Species in Electrochemical, Photochemical, and Photoelectrochemical Water Oxidation Reactions in Phosphate Electrolytes. *Chem - A Eur J* n/a–n/a. doi: 10.1002/chem.201502428
15. Dimitrou K, Brown AD, Christou G, et al. (2001) Mixed-valence, tetranuclear cobalt(III,IV) complexes: preparation and properties of [Co<sub>4</sub>O<sub>4</sub>(O<sub>2</sub>CR)<sub>2</sub>(bpy)<sub>4</sub>]<sup>3+</sup> salts. *Chem Commun* 4:1284–1285. doi: 10.1039/b102008k

16. Evangelisti F, Guettinger R, More R, et al. (2013) Closer to Photosystem II : A Co O Cubane Catalyst with Flexible Ligand Architecture Closer to Photosystem II : A Co 4 O 4 Cubane Catalyst with Flexible Ligand Architecture. *J Am Chem Soc* 135:18734–18737. doi: 10.1021/ja4098302
17. McCool NS, Robinson DM, Sheats JE, Dismukes GC (2011) A Co<sub>4</sub>O<sub>4</sub> cubane water oxidation catalyst inspired by photosynthesis. *J Am Chem Soc* 133:11446–11449. doi: 10.1021/ja203877y
18. Berardi S, La Ganga G, Natali M, et al. (2012) Photocatalytic water oxidation: tuning light-induced electron transfer by molecular Co<sub>4</sub>O<sub>4</sub> cores. *J Am Chem Soc* 134:11104–7. doi: 10.1021/ja303951z
19. Chakrabarty R, Bora SJ, Das BK (2007) Synthesis , Structure , Spectral and Electrochemical Properties , and Catalytic Use of Cobalt ( III ) – Oxo Cubane Clusters. *Polyhedron* 26:9450–9462.
20. Najafpour MM, Rahimi F, Aro E-M, et al. (2012) Nano-sized manganese oxides as biomimetic catalysts for water oxidation in artificial photosynthesis: a review. *J R Soc Interface* 9:2383–2395. doi: 10.1098/rsif.2012.0412
21. Tantillo DJ, Chen J, Houk KN (1998) Theozymes and compuzymes: theoretical models for biological catalysis. *Curr Opin Chem Biol* 2:743–750. doi: 10.1016/S1367-5931(98)80112-9
22. Siegel JB, Zanghellini A, Lovick HM, et al. (2010) Computational Design of an Enzyme Catalyst for a stereoselective Bimolecular Diels-Alder Reaction. *Science* 328:105:1–6.
23. Röthlisberger D, Khersonsky O, Wollacott AM, et al. (2008) Kemp elimination catalysts by computational enzyme design. *Nature* 453:190–195. doi: 10.1038/nature06879
24. Jiang L, Althoff E a, Clemente FR, et al. (2008) De novo computational design of retro-aldol enzymes. *Science* 319:1387–1391. doi: 10.1126/science.1152692
25. Bradley P, Misura KMS, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309:1868–1871. doi: 10.1126/science.1113801
26. Mandell DJ, Kortemme T (2009) Backbone flexibility in computational protein design. *Curr Opin Biotechnol* 20:420–428. doi: 10.1016/j.copbio.2009.07.006
27. Mandell DJ, Coutsiadis E a, Kortemme T (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods* 6:551–552. doi: 10.1038/nmeth0809-551

## **Chapter 3:      Benchmarking a computational design method for the incorporation of metal ion-binding sites at symmetric protein interfaces**

### **3.1 Preface**

A version of this chapter has been published in Protein Science and is formatted in the journal style.

### **3.2 Abstract**

The design of novel metal-ion binding sites along symmetric axes in protein oligomers could provide new avenues for metalloenzyme design, construction of protein-based nanomaterials and novel ion transport systems. Here, we describe a computational design method, symmetric protein recursive ion-cofactor sampling (SyPRIS), for locating constellations of backbone positions within oligomeric protein structures that are capable of supporting desired symmetrically coordinated metal ion(s) chelated by sidechains (chelant model). Using SyPRIS on a curated benchmark set of protein structures with symmetric metal binding sites, we found high recovery of native metal coordinating rotamers: in 65 of the 67 (97.0%) cases, native rotamers featured in the best scoring model while in the remaining cases native rotamers were found within the top three scoring models. In a second test, chelant models were crossmatched against protein structures with identical cyclic symmetry. In addition to recovering all native placements, 10.4% (8939/86013) of the non-native placements, had acceptable geometric compatibility scores. Discrimination between native and non-native metal site placements was further enhanced upon constrained energy minimization using the Rosetta energy

function. Upon sequence design of the surrounding first-shell residues, we found further stabilization of native placements and a small but significant (1.7%) number of non-native placement-based sites with favorable Rosetta energies, indicating their designability in existing protein interfaces. The generality of the SyPRIS approach allows the design of novel symmetric metal sites, including non-natural amino acid sidechains, and should enable the predictive incorporation of a variety of metal-containing cofactors at symmetric protein interfaces.

### **3.3 Introduction**

In nature, 30% of proteins are believed to contain metal binding sites,<sup>1</sup> which exhibit a diverse array of functional and structural utility.<sup>2-5</sup> Metal sites are observed across all protein topologies and microenvironments: buried within the protein core, at or near the surface, or positioned at protein interfaces.<sup>6, 7</sup> In some cases, metals can be coordinated along axes of symmetry of homo-oligomeric proteins by symmetry-related sidechains from the individual subunits.<sup>8</sup> These metal sites at homo-oligomeric interfaces increase protein thermodynamic stability, and can play functional roles as well as induce supramolecular assembly formation in engineered<sup>6, 9, 10</sup>, and natural proteins, for example, ferritin<sup>11</sup>.

Computational metalloprotein design has applications in industrial biocatalysis, pharmaceutical production, and biomaterial engineering. However, it remains a challenge as the accurate design of metal–protein interactions require precise coordination geometry, and therefore, precise placement of protein sidechains, typically in a relatively hydrophobic environment. Repurposing existing metal binding sites has been feasible,<sup>12</sup>

and computational strategies such as METALSEARCH<sup>13</sup> and Dezymer<sup>14</sup> have been used to identify locations for novel metal binding sites with a tetrahedral coordination geometry.<sup>15-17</sup> Rational design of the geometrically well-defined helix bundle topology made possible the incorporation of multinuclear metal clusters and cofactors.<sup>18-26</sup> In these and other design efforts, however, crystallographic structural validation was not reported.<sup>24, 26-33</sup> Recently, a computational algorithm, URANTEIN, was developed for designing uranyl-binding sites. Crystal structures and design models show high agreement in the backbone conformation; however, 40% of the binding residues showed altered conformations.<sup>34</sup> The Rosetta macromolecular modeling suite<sup>35,36</sup> has made possible the general and rapid search of protein topologies for novel metal chelation. RosettaMatch was used to incorporate mononuclear Zn binding sites including those designed with unnatural amino acids.<sup>37,38</sup> In each case, the desired coordination (model) did not match the crystal structure with atomic-level accuracy, which is likely a design requirement for engineering efficient catalysis.<sup>39</sup> Near atomic-level accuracy has recently been achieved by designing with high affinity chelating ligands and utilizing symmetric metal chelation. For example, a crystallographic structure validated the model of a tris-2,2'-bispyridine-Fe<sup>2+</sup> complex incorporated along an oligomeric axis of symmetry.<sup>40</sup> Thus, despite wide-ranging successes, new methods that increase the accuracy of metal coordinating sidechain placement are needed in computational metalloprotein design.

In this article, we report the development of a rapid computational search and design method aimed at incorporating mononuclear, multinuclear, and cofactor bound metal sites along symmetry axes of homo-oligomeric proteins. The method, that we call Symmetric Protein Recursive Ion-cofactor Sampling (SyPRIS), utilizes an inverse

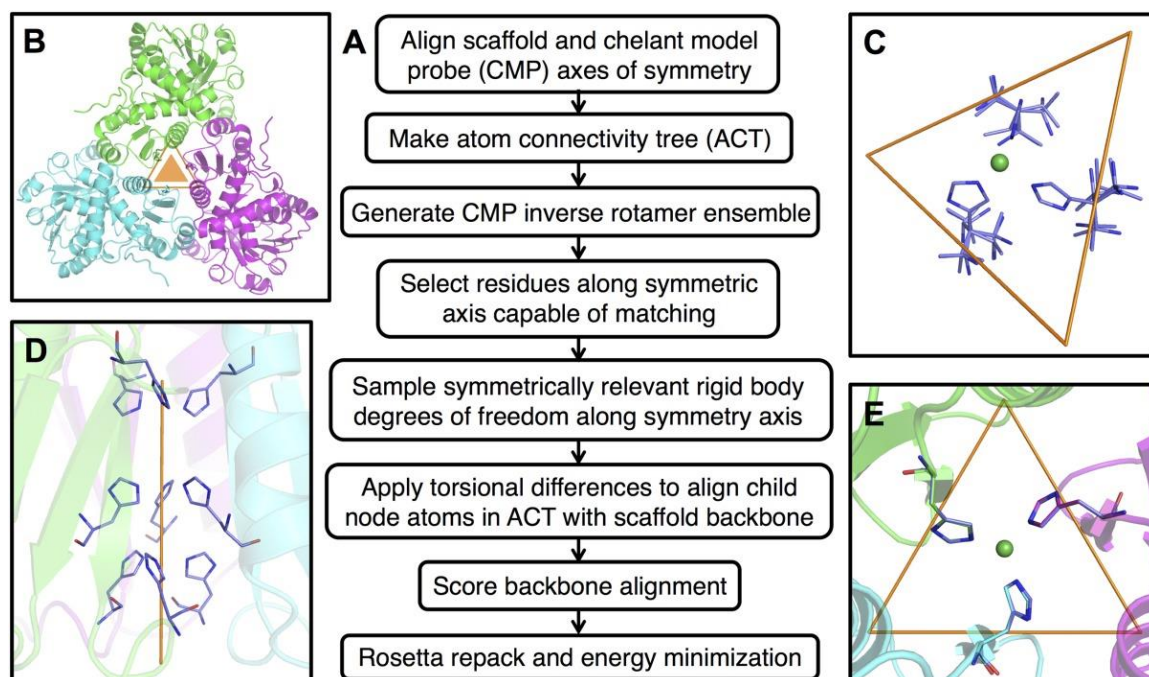
rotamer search algorithm and Rosetta energy optimization. A model of symmetrically placed, unconnected amino acids with a given coordination geometry (a chelant model probe, or CMP) is positioned at geometrically compatible locations along an oligomeric axis of symmetry by sampling the symmetry-allowed rigid body and rotameric degrees of freedom. A CMP can be generated *de novo* similar to a theozyme<sup>41</sup>, or derived from an existing protein structure. Once a metal binding site are located with SyPRIS, rotamer optimization and sequence design is performed using Rosetta. To determine the prediction accuracy of SyPRIS, we curated a benchmark set of 67 oligomeric metalloprotein structures from the Protein Data Bank.<sup>42</sup> Each structure contains a metal center along a cyclic symmetric axis. CMPs were derived from the native structures by extracting the metal center and chelating residues. We first recapitulate the native residue position and rotamer identities of the native proteins. We then crossmatched the CMP and scaffold libraries with identical symmetry and performed energy minimization to test discrimination between native and non-native matches. We show that SyPRIS is capable of recapitulating, with high accuracy, the binding geometry of all natural metalloproteins tested.

## 3.4 Results

### 3.4.1 Summary of algorithm and workflow

We have developed a method for locating protein backbone geometries suitable for symmetric metal chelation (Fig. 1). First, the geometric centers and the relevant axes of symmetry of the protein scaffold and the chelant model probe (CMP) are aligned [Fig. 1

(A)]. A map of covalently bonded atoms, termed the atom connectivity tree (ACT), is created from the CMP before sampling. A CMP has an asymmetric unit—one or more amino acid residues with fixed metal or ligand coordination geometry—which is replicated by rotation about one or more symmetric axes to obtain the final metal-containing probe [Fig. 1(B)]. An ensemble of inverse rotameric CMP states is generated using the ACT and a backbone-independent rotamer library. We then sample the symmetrically relevant rigid body degrees of freedom using each member of the CMP ensemble as a probe. In this way we locate compatible backbone atom positions along target protein symmetry axes while avoiding unnecessary clashes [Fig. 1(C)]. To evaluate backbone overlap between the scaffold and the placed CMP, we developed a geometric metric (SyPRIS score) that includes both the proximity and geometric directionality of the matched backbone atoms (see "Methods"). Matches with acceptable SyPRIS scores are grafted into the scaffold, followed by RosettaDesign sequence/structure optimization.



**Figure 8-3. Figure 1. The computational workflow of the SyPRIS algorithm.**

The chelant model probe (CMP), three histidines coordinating a metal center about the axis of symmetry, is used to locate compatible backbone geometries along a trimeric symmetry axis of the native trimeric scaffold protein. **(A)** Flowchart of steps from placing CMP to design and energy minimization. **(B)** The symmetry of the CMP and scaffold are determined and aligned. **(C)** An ensemble of base rotamers is generated using the atom connectivity tree and rotamer library. **(D)** The CMP is used to probe along the axis of symmetry for compatible scaffold backbone locations. **(E)** A native match, CMP, and corresponding scaffold residues shown in sticks.

### 3.4.2 Identification of native metal binding sites with SyPRIS

To parameterize the SyPRIS algorithm, we first attempted to recapitulate symmetric metal binding sites found in native proteins (Fig. 2). To this end, we used SyPRIS to determine the geometries of sites that confer metal binding within a database of 67 protein crystal structures featuring cyclic symmetry. The database we curated is composed of 44 dimers, 22 trimers, and one tetramer with metal-binding sites along symmetry axes (Supporting Information Table S1).

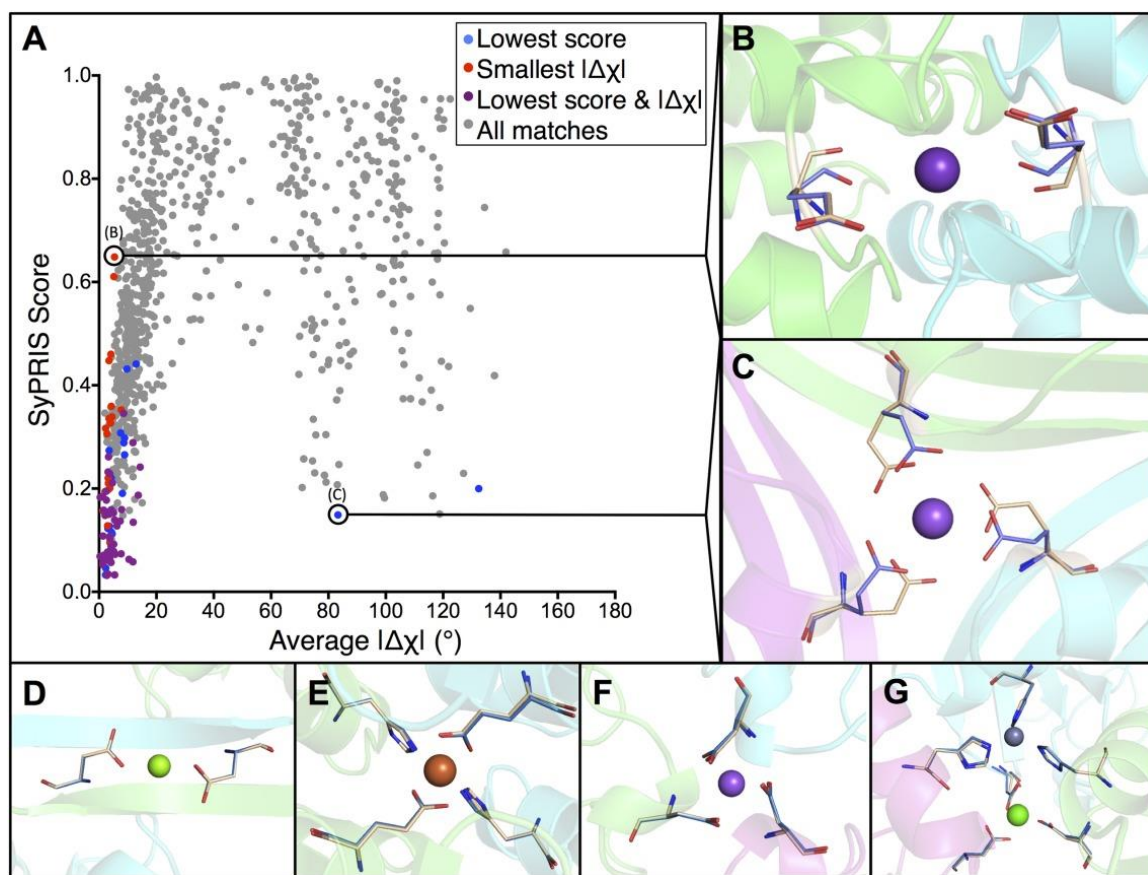


Figure 9-3. Figure 2. Native SyPRIS match results for a benchmark of 67 native metal chelating homo-oligomeric proteins.

(A) The plot of SyPRIS score versus the absolute difference of the metal-chelating sidechain Torsion angles (ADTA) from CMP to wildtype residue. We highlight the 67 lowest SyPRIS scoring matches (blue) and the 67 matches with the smallest ADTA (red). Matches where the lowest SyPRIS score corresponds to the lowest ADTA are also indicated (purple). (B) Example of outlier smallest-ADTA match with a SyPRIS score above 0.5. (C) Example of outlier lowest-SyPRIS-score match with ADTA above  $80^\circ$  illustrating an alternative rotamer that satisfies the metal coordination geometry along the symmetric axis. Examples of accurate matches: (D) Dimeric mononuclear single-residue match. (E) Dimeric mononuclear multi-residue match. (F) Trimeric mononuclear single-residue match. (G) Trimeric dinuclear multi-residue match. Protein scaffolds are shown colored by chain (green, cyan, and magenta), target native residue locations are shown in sticks (wheat). Corresponding CMP matches are shown as sticks (skyblue).

We first performed SyPRIS-based native CMP placements at native scaffold residue positions and examined how well the SyPRIS score could recapitulate the native rotamer.

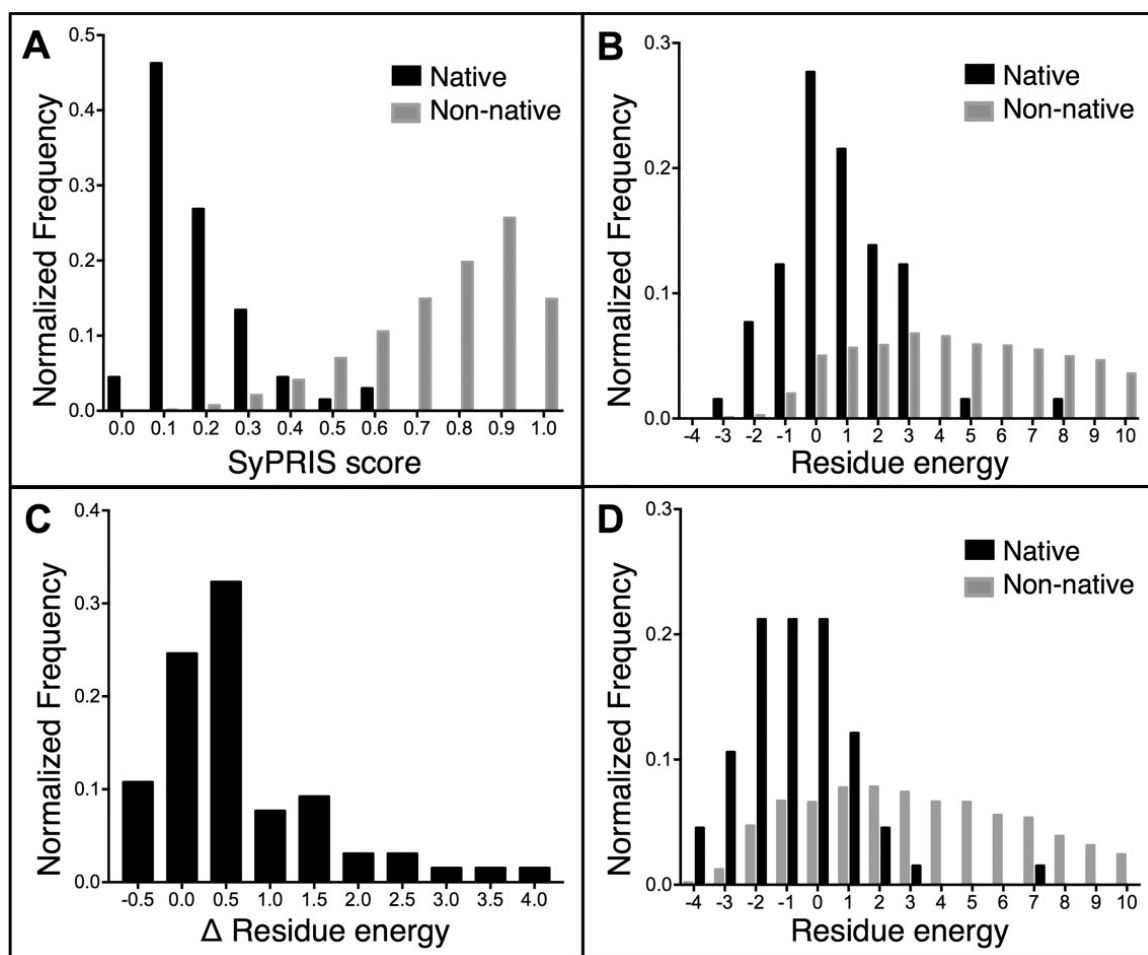
The SyPRIS score was compared with the average absolute difference of the metal-chelating sidechain torsion angles (ADTA) between the predicted CMP match and the corresponding native scaffold rotamers [Fig. 2(A)]. ADTA is a direct measure of how well the CMP matches the native rotamer; a value less than  $20^\circ$  was considered native. The SyPRIS score calculates the CMP-to-scaffold backbone atom overlap and has a value of zero for a perfect backbone alignment. 48 of the 67 (71.6%) of the matches with the smallest ADTA were also the lowest SyPRIS scoring matches. Within a SyPRIS score threshold of 0.5, the smallest ADTA match was found in 65 of the 67 (97%) matches. For the remaining two protein outliers, the CMPs are composed of sidechains with three rotatable sidechain torsional angles [Fig. 2(B)]. The combinatorics of discretization makes the absence of native rotamer more likely for longer sidechains. Therefore, these two outliers are likely caused by the discretization of the rotamer sampling for sidechains with three or greater torsion angles.

Notably, of the 67 scaffolds tested, 65 scaffolds (97%) have CMP matches for which the lowest SyPRIS score corresponds to an ADTA less than  $20^\circ$  [Fig. 2(A)]. In both outlier cases, an alternative rotamer and metal site position along the symmetry axis satisfied the CMP metal-binding geometry with substantial backbone overlap to the native residue location [Fig. 2(C)]. Based on these results, we chose to accept all matches with a SyPRIS score of less than 0.5 for further optimization. Indeed, a native-like CMP match existed (ADTA  $<20^\circ$ ) within the top three SyPRIS scores ( $<0.5$ ) for all 67 proteins [Fig. 2(D–G)]. Thus, we conclude that SyPRIS recapitulates placements of metal chelating sidechain rotamers well, and that the backbone-independent library used was sufficiently populated with chelation-competent rotamers.

### 3.4.3 Location of metal binding sites from cross-CMP matching with SyPRIS

The predictive capability of small molecule protein docking and enzyme design methods has been tested by cross-docking/placement benchmarks, in which a known ligand binding or active site is grafted into proteins of the same family.<sup>36, 43-45</sup> The goal is to develop a sampling approach that will discriminate between native and non-native placements of the small molecule/binding partner based on geometric and energetic criteria.

To investigate if our approach can similarly discriminate between native and non-native CMP placements, we used SyPRIS to cross match the CMP and benchmark libraries (with identical symmetry) to locate all geometrically compatible placements. We compared the distribution of SyPRIS scores over 86,013 non-native CMP placements to the distribution of scores for the 66 native residues CMP placements [Fig. 3(A)]. We observed that as SyPRIS score increases, the frequency of the native matches decays while the frequency of non-native matches grows rapidly. Of the 86,013 non-native CMP matches, 8938 (10.39%) matches scored within a SyPRIS score threshold of 0.5. The non-native matches correspond to CMP placements with alternative sequences or scaffold identities. Thus, as a first step, a SyPRIS score threshold-based filtering recapitulates all native matches and produces non-native matches with native-like geometric compatibility scores.



**Figure 10-3. Figure 3.** Comparison of native and non-native matches with SyPRIS score and Rosetta residue energy.

(A) SyPRIS score distribution for the 66 native matches (black) and the 86,013 non-native matches (gray). (B) Rosetta residue energy distribution after energy-based minimization for the 66 natives matches (black) and the 8,938 non-native matches (gray) with a SyPRIS score threshold of 0.5. (C) The residue energy differences between the native scaffold and SyPRIS-placed models. (D) Rosetta residue energy distribution for SyPRIS-placed chelating residues after RosettaDesign. The design was performed on all matches identified in B.

### 3.4.4 Energy minimization improves discrimination

In the next step, aimed at discriminating native from non-native placements, we applied a Rosetta-based energy optimization protocol to the 8938 non-native and 66 native CMP matches that were chosen based on their SyPRIS scores in the first step. For each matched position and rotamer, we grafted the SyPRIS-identified rotamer on the scaffold

backbone, and performed sidechain repacking followed by gradient-based minimization.<sup>46</sup> Atom coordinate constraints were applied to the SyPRIS-identified CMP atom locations to enforce the desired coordination geometry. We evaluated the energy optimization with the absolute residue energy of the placed chelating residues and found the resulting average native and non-native residue energies to be 0.8 and 4.3 Rosetta energy units (REUs), respectively [Fig. 3(B)]. Thus, illustrating an energy gap separates the native and non-native placements. A total of 60 native (out of 67) and 2004 non-native (out of ~86,000) structures score less than 3.0 REUs [Fig. 3(B)], indicating that Rosetta energy minimization dramatically reduces the number of non-native matches.

To investigate if positive (>0.0 Rosetta Energy Units) residue energies that were frequently observed for the SyPRIS-placed metal-chelating residues are a feature of metal-binding residues in proteins or an artifact of our placement algorithm, we performed the same minimization protocol on the native scaffold (before CMP replacement) and compared the difference in residue energies to the SyPRIS-placed variants [Fig. 3(C)]. Overall, we found within this benchmark that positive metal-chelating residue energy is common in the native metal cheating residues and is caused by strained rotamer identities. However, seven postminimized native models were calculated to be more stable than their SyPRIS-placed counterparts by greater than 2.0 REUs indicating that SyPRIS-based placement had led to these unphysical high energies in the models. We found that these differences could be attributed to symmetry-related mismatches between the CMP (derived from the crystal structure) and the Rosetta-idealized scaffolds. During Rosetta Symmetry modeling,<sup>47</sup> the scaffold homo-oligomeric protein is "idealized" in such a way that the same symmetric transform relates every atom

in each protomer. As crystallographic structures can have minor deviations from perfect symmetry, this symmetric idealization can, in some cases, lead to changes in protomer placement. In all seven cases, translations of 1 Å or greater (away from the symmetric axis) were introduced while idealizing the crystallographic structure with Rosetta Symmetry. Translation away from the symmetry axis during Rosetta-based idealization in these seven cases led to high coordinate-constraint penalties on the CMP and ultimately resulted in unfavorable residue energies.

To ensure that the few<sup>7</sup> unphysical high residue energies were due to the minor changes in protein symmetry upon idealization, we performed the same protocol with a CMP derived from the idealized scaffold. When using the idealized version of the CMP, we resolved the unphysical high energies in all seven cases (data not shown). Despite the improved recovery of native when using the idealized CMP, we chose the nonidealized CMP to preserve crystallographic coordination geometries and to avoid conformational memory bias when sampling. However, symmetric idealization of the scaffold remains a necessity for two reasons: enhanced accuracy of rigid body sampling along the symmetry axes and to access the Rosetta Symmetry design protocols (symmetric packing/design) which both require and imposes symmetric idealization.

These results show that the combination of the SyPRIS-based geometric compatibility and energy evaluation (postminimization) combine to discriminate native from non-native matches. Minor deviations in symmetry can, in some cases, lead to artifactually high residue energies for metal-chelating residues, highlighting the sensitivity of energetics to minor changes in symmetry.

### 3.4.5 Design of sites identified using SyPRIS

Having demonstrated that the combination of geometric match compatibility and energy minimization/evaluation can produce native-like symmetric metal binding sites, we asked if RosettaDesign could optimize the native residue energies and obtain native-like energies for non-native matches. The design was performed following standard enzyme design protocols<sup>46</sup> in Rosetta (RosettaScripts files for the repacking and design protocols are provided in the Supplementary Information). The native and average non-native energies decreased from 0.8 and 4.2 [after minimization, Fig. 3(B)] to  $-0.7$  and  $3.0$  REUs, respectively [Fig. 3(D)]. While the native and non-native discrimination was maintained, 1469 designed non-native metal binding sites with residue energies below  $0.0$  REUs were detected. These results demonstrate that a SyPRIS-based design protocol can recover all native metal-binding sites with favorable energies and generate candidate novel designs with similar energy values.

### 3.4.6 Cross-matching results for similar CMPs

The coordination geometry of ligands greatly contributes to the catalytic function and efficiency of metal sites in proteins. For example, the same sidechain functional groups can bind a given metal ion in tetrahedral, octahedral, or trigonal bipyramidal geometries and serve dramatically different functions. To determine the sensitivity of SyPRIS to small variations in coordination geometry for a given set of chelating amino acid residue types, we selected a subset of eight trimeric scaffolds with native CMPs composed of histidine residues and various coordination geometries. All but one CMP, in the subset, coordinate the metal ion with the N $\epsilon$ 2 nitrogen atom in the imidazole ring. We performed

CMP alignment, pairing the N $\epsilon$ 2, N $\delta$ 1, C $\gamma$ , and C $\beta$  atoms of the sidechain across the CMP subset library. The RMSD of the CMP alignments provides a measure of coordination geometry similarity and range from 0.3 Å (2VRS and 3BF3) to 1.9 Å (1BCH and 1ULI). To avoid artifacts previously encountered due to RosettaSymmetry-based idealization, we derived the CMP geometries from the idealized structures. We then crossmatched the subset of CMPs and found that of the 56 non-native CMP to scaffold matching pairs, 48 have SyPRIS scores greater than 0.5—indicating their geometric incompatibility.

We performed energy optimization of the eight matches with a SyPRIS score of less than 0.5 and compared the resulting Rosetta residue energies (Fig. 4). In the native CMP to native scaffold matches, the absolute Rosetta residue energy ranged from 0.2 to 2.7 REUs. Three of the eight had Rosetta residue energies over 5.0 REUs indicating their incompatibility; as expected, these three also had the highest coordination geometry RMSDs [Fig. 4(A,B)]. The remaining 5 CMP-scaffold match pair residue energies ranged from 0.1 to 2.6 REUs—comparable to the native matches. Additionally, the five non-native matches with native-like energies had the smallest coordination geometry RMSDs (<0.6), indicating their structural similarity with the native CMP [Fig. 4(C)]. These results suggest that the combination of a SyPRIS score match threshold and a subsequent Rosetta residue energy threshold can be used to discriminate robustly between differences in metal coordination geometry during placement and design.

<b>A</b>		<b>Scaffold Protein</b>							
		<b>3BF3</b>	<b>2W37</b>	<b>2VRS</b>	<b>2J4J</b>	<b>1ZEI</b>	<b>1ULI</b>	<b>1EF8</b>	<b>1BCH</b>
<b>1BCH</b>	RMSD (Å)	1.8	1.7	1.9	1.7	1.8	1.9	1.8	0.0
	SyPRIS score	-	-	-	0.4	-	-	-	0.2
	Residue energy	-	-	-	5.2	-	-	-	0.6
<b>1EF8</b>	RMSD (Å)	0.5	0.6	0.6	0.5	1.4	1.4	0.0	1.8
	SyPRIS score	0.9	0.3	-	0.3	-	-	0.2	0.6
	Residue energy	-	1.5	-	2.4	-	-	2.7	-
<b>1ULI</b>	RMSD (Å)	1.4	1.6	1.3	1.7	0.9	0.0	1.4	1.9
	SyPRIS score	-	0.8	1.0	0.9	-	0.1	-	1.0
	Residue energy	-	-	-	-	-	0.4	-	-
<b>1ZEI</b>	RMSD (Å)	1.0	1.2	0.9	1.4	0.0	0.9	1.4	1.8
	SyPRIS score	-	1.0	-	-	0.4	0.4	-	-
	Residue energy	-	-	-	-	1.5	21.7	-	-
<b>2J4J</b>	RMSD (Å)	0.6	0.5	0.8	0.0	1.4	1.7	0.5	1.7
	SyPRIS score	-	0.6	-	0.2	-	-	0.3	0.6
	Residue energy	-	-	-	1.2	-	-	2.6	-
<b>2VRS</b>	RMSD (Å)	0.3	0.7	0.0	0.8	0.9	1.3	0.6	1.9
	SyPRIS score	0.4	0.5	0.1	0.9	0.8	0.7	-	-
	Residue energy	0.1	12.8	2.2	-	-	-	-	-
<b>2W37</b>	RMSD (Å)	0.4	0.0	0.7	0.5	1.2	1.6	0.6	1.7
	SyPRIS score	-	0.1	-	0.6	-	-	0.4	-
	Residue energy	-	0.9	-	-	-	-	2.6	-
<b>3BF3</b>	RMSD (Å)	0.0	0.4	0.3	0.6	1.0	1.4	0.5	1.8
	SyPRIS score	0.2	0.9	0.7	-	0.9	0.8	0.9	-
	Residue energy	0.2	-	-	-	-	-	-	-

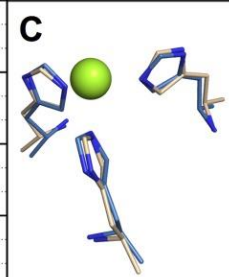
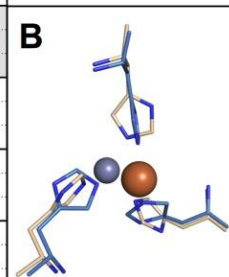


Figure 11-3. Figure 4. Cross-match results for symmetric CMPs composed of three histidine residues into trimeric scaffolds at native metal-binding histidine locations.

(A) The degree of alignment of CMP pairs (coordination geometry) is measured as the RMSD across the C $\beta$ , C $\gamma$ , N $\delta$ 1, and N $\epsilon$ 2 atoms of two histidine residues when one histidine sidechain is used for alignment. SyPRIS score values less than the 0.5 threshold value are highlighted in red. Rosetta energy-based minimization was performed on matches that scored within the SyPRIS threshold; Rosetta residue energies, postminimization, are indicated. (B) 1ZEI (CMP) aligned to 1ULI (native scaffold residues). This match reports the highest postminimization residue energy in A. (C) 2VRS (CMP) aligned to 3BF3 (native scaffold residues). This match reports the lowest postminimization residue energy in A.

### 3.5 Discussion

We describe a computational method, SyPRIS, and a benchmarking test for the design of symmetric metal-binding sites at homo-oligomeric protein interfaces. In our benchmarking test, the method is capable of determining all native rotamer identities of the 67 tested naturally-occurring symmetric metal sites within the top 3 scored matches. Energy-based optimization of the resulting matches led to robust discrimination between native and non-native placements. Upon sequence design, a small but significant number

of potential new metal-binding sites were predicted with Rosetta energies comparable to their wild type structures, highlighting the designability of homo-oligomeric interfaces for novel metal chelation properties. While the benchmark test consisted of *C*-symmetric homo-oligomeric metalloproteins with canonical amino acid-based chelation, the SyPRIS algorithm is not limited to *C*-symmetric scaffolds, canonical amino acids, or metal chelation. Any symmetry observed in the protein databank (e.g., *D*-symmetric proteins), noncanonical amino acids, nonmetal-containing cofactors, and transition state models with minor asymmetries could be placed using the same approach. As long as the amino acid residues incorporated are symmetric, small asymmetries in bound ligands are tolerated.

Apart from providing novel microenvironments for the design of functional sites, homo-oligomeric interfaces offer a key advantage for the design of novel metalloproteins. Most metal-binding sites require three or more chelating sidechains. In asymmetric metalloprotein design, each coordinating residue is required to adopt a precise geometry compatible with metal chelation. Every metal chelating residue has a probability of adopting an imprecise geometry due to the lack of (or a small) energy gap between designed and alternative conformations.<sup>38</sup> Therefore, the probability of adopting imprecise geometry increases with the number of independent residues used in the design. However, if the metal-binding site is itself symmetric (e.g., a trimer interface), the same residue conformation is likely to be replicated across all subunits. In this way, we need only consider alternative conformations of a single chelating residue. Protein symmetry coupled with metal chelation can expand the energy gap between desired binding-competent and alternative conformations.<sup>48</sup> Thus, inherent symmetry increases

the likelihood that designed residues adopt intended binding-competent rotameric identities. This is one explanation for the notable successes achieved in metal and cofactor incorporation in symmetric helical bundles,<sup>18-26, 29</sup> and for the recent design of a tris-2,2'-bipyridine-based metalloprotein.<sup>40</sup>

The SyPRIS algorithm uses an inverse rotamer library, not unlike previously described algorithms,<sup>13, 14, 36</sup> to locate viable residue positions for grafting metal-binding sites at oligomeric interfaces. With inverse rotamer sampling, speed scales at least linearly with the number of rotameric states sampled. The algorithm is less efficient than alternative algorithms that utilize geometric hashing.<sup>36</sup> However, constraints offered by symmetry limit the size of the search space for placement, rendering speed not a significant impediment in our usage. In principle, similar placement can be obtained using geometric hashing-based algorithms such as RosettaMatch, but RosettaMatch will produce all symmetric and asymmetric matches that satisfy the coordination geometry of metal binding. Workaround solutions are necessary to locate the matches that satisfy the constraint parameters symmetrically. Additionally, the algorithms are complementary in their approach. RosettaMatch grows the sidechains directly from the backbone to evaluate the coordination geometry, while SyPRIS maintains the inherent coordination geometry and evaluates the geometric compatibility for a given scaffold backbone by measuring the inverse rotamer backbone-atom overlap. A combination of geometric hashing and SyPRIS-based placement algorithms could be useful for problems involving a large number of chelating residues and is the focus of future algorithm development in our group. We have also successfully utilized SyPRIS to design novel metal ion-

coordinating binding sites with noncanonical amino acids, and these will be reported elsewhere.

Many challenges remain in metalloprotein design, such as: designing complex catalytic function and determining and modulating the metal-ion preference of a designed binding site. These challenges are particularly difficult to solve when modeling metal ion interactions with an empirical force field where metal-binding interactions are treated as constraints/restraints. One possibility for future efforts may be to use more accurate quantum mechanical simulation methods to evaluate and refine metal ion binding (and preference) generated by computationally inexpensive algorithms such as the one described here.

## **3.6 Methods**

### **3.6.1 Preparing the benchmark native scaffold and CMP databases**

A library comprised of symmetric non-redundant homo-oligomerically coordinating metalloproteins served as a diverse benchmark for this study. The benchmark was first curated from the Mespeus database.<sup>49</sup> Protein Data Bank (PDB)<sup>42</sup> entries within the Mespeus database reported as bound to a metal by the same residue type on more than one chain were selected for the benchmark. 84 protein structures deemed nonredundant were subject to preparatory scripts (Supporting Information) to extract the CMP, create Rosetta symmetry files, and generate idealized homo-oligomeric Fast Relaxed<sup>46</sup> protein structures. CMPs were extracted from the original wildtype crystal structure files. We aligned and calculated the RMSD of the native wild type structure against the native FastRelaxed structure and omitted 17 structures from the benchmark found to have asymmetries in the wildtype ( $\text{RMSD} > 1.0 \text{ \AA}$ )—creating a final library of 67 scaffolds. Both the complete benchmark library and the full set of homo-oligomeric PDB codes from the Mespeus database can be found in the supplementary information (Supporting Information Tables S1 and S2).

### 3.6.2 Symmetric protein recursive ion sampler (SyPRIS) package

As described previously,<sup>50</sup> SyPRIS is a Python-based program that recursively samples all CMP rigid body and rotamer degrees of freedom to locate sidechain constellations compatible with the backbone positions along a symmetry axis of a target protein scaffold suitable for grafting CMPs. A detailed step-by-step description of the SyPRIS matching algorithm is presented in Supporting Information. Briefly, the algorithm consists of six steps:

*Step 1:* Read a given backbone-dependent rotamer library. A backbone independent rotamer dictionary is constructed from a backbone dependent library where rotamer identities are averaged over all backbone dihedral combinations. The identity is stored as a key whose values are the rotamer torsion angles and respective standard deviations.

*Step 2:* Construction of the CMP object. In the CMP object, SyPRIS identifies substructures based on Cartesian coordinates: metal center, chelating amino acid, and additional cofactor atoms. Chelating amino acids are further dissected and organized into fulcrum atom, and continuously and discretely sampled torsion angles within the sidechain [Fig. 5(A)].

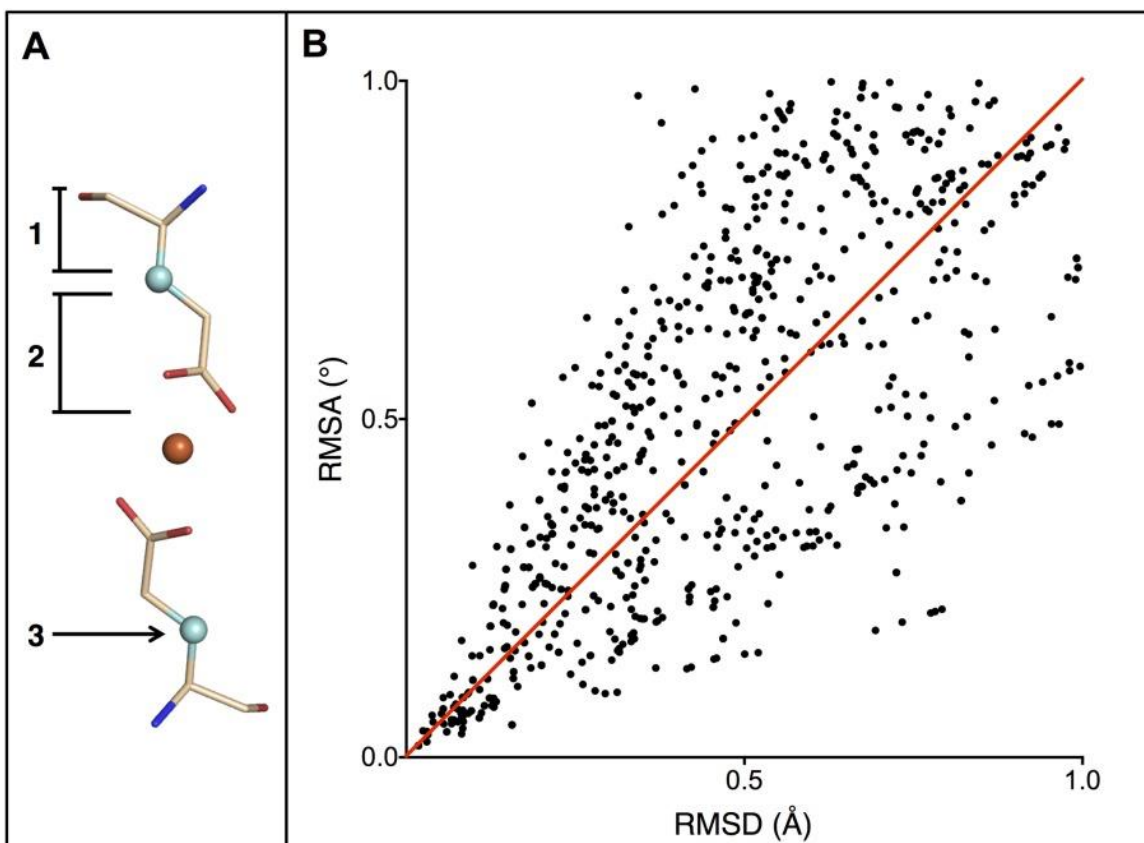


Figure 12-3. Figure 5. SyPRIS-related terms and parameters.

(A) **1**: Atom positions/torsions sampled continuously (rotate freely within standard deviation). **2**: Atom positions/torsions sampled discretely (restricted to base rotamers). **3**: The C $\beta$  fulcrum atom that separates continuously and discretely sampled torsion angles. (B) RMSA versus RMSD plot of native residue match set. The red line is the preferred target slope for the data (not a trend line).

*Step 3*: Generation of CMP base rotamer ensemble. The CMP is aligned to the axes of symmetry of the target protein. Using the rotamer dictionary described in *Step 1*, CMP torsions are rotated to match each value for a given key in the dictionary—creating a base rotamer. Base rotamers are stored within the object. Additional rotamers can be generated with the standard deviation multiplier.

*Step 4*: Locate potential residue match positions along the axis of symmetry. For each base rotamer generated in Step 3, the fulcrum atom radius is used to determine the

candidate match-competent residues along the symmetry axis. Candidate residues located outside of the radial distance  $\pm 1$  Å are not sampled. We sample the C-symmetric rigid body degrees of freedom by translating along the c-symmetric axis to each candidate residue and rotating the CMP such that the fulcrum atom comes within a minimum distance of the corresponding scaffold main chain atom [Fig. 1(C)].

*Step 5: Continuous  $\chi$  torsion sampling.* Starting from the base rotamer torsion angle, the continuously sampled torsion angles (backbone atom positions) are rotated until alignment to the scaffold backbone or until the maximum value within the torsion standard deviation. The CMP is then subject to a clash check with the scaffold backbone to discard clashing matches.

*Step 6: Scoring match outputs.* To score a placed CMP for geometric compatibility, we generated a term that considers the proximity and the directionality of the aligned CMP to scaffold backbone atoms. RMSD serves as a measure for proximity but to determine the directionality of the connected atom vectors, we developed a metric that we call root-mean squared angle (RMSA).

$$\text{RMSA} = \frac{6}{\pi} \sqrt{\left( \sum_{i=0}^n \theta_i \right)^2 / n} \quad (1)$$

where  $\theta$  is the angle difference between each CMP vector (bond) and the corresponding vector in the scaffold. We chose the functional form in [1](#)) for this term as it scales linearly with RMSD from 0 to 1 and becomes less correlated as each term approaches 1 [Fig.

[5\(B\)](#)]. The SyPRIS score is the average of RMSA and RMSD summed over all peptide backbone bonds and atomic positions, respectively.

### 3.6.3 Native match recapitulation and rotamer identification

A wrapper application (see Supporting Information Methods) was used to sample each CMP within the corresponding scaffold. Other adjustable parameters (the identity of the fulcrum atom, RMSA, clash radius) were set to default values. Each output match is stored according to the PDB identity of the scaffold and CMP (the latter for benchmarking purposes), matched residue index, matching base rotamer indices, associated  $\chi$  torsion angles, RMSD, RMSA, and the calculated SyPRIS score.

### 3.6.4 Cross-match discrimination and design of native and non-native CMPs

Energy-based discrimination was performed with two similar protocols: minimization-only and design. SyPRIS settings for these experiments were identical to the native only experiment except that we crossmatched CMP and scaffold libraries of similar symmetry. CMP matches that passed a 0.5 SyPRIS score threshold were grafted back into their respective scaffolds via residue replacement at the matched backbone locations using PyRosetta.[51](#) For repacking simulations, sidechains within 8 Å of the metal center were allowed to change their rotameric configuration followed by gradient-based energy minimization. For design simulations, residues within 8 Å of the CMP match site sampled different residue types as well as rotamers. Native residue identities were up-weighted by a factor of 1.5. Atomic coordinate constraints were placed on the metal-binding atom positions of the pregrafted CMP (derived from the crystal structure). While

repacking/minimizing, the coordinate constraints would impose an energy penalty if the replaced residues deviated from the intended metal coordination. The energy optimization protocol we used is similar to the Fast Relax protocol<sup>46</sup> with fixed backbones. RosettaScripts files, and the command-lines used for repacking and design simulations are provided in the Supporting Information.

### 3.6.5 Geometric coordination deviation impact across tris-his CMPs

To determine the sensitivity of SyPRIS for matching non-native sites with CMPs that vary only in coordination geometry, we curated a smaller subset library from the main benchmark library. The subset library was comprised of trimeric scaffolds that coordinate a metal along the symmetric axis with a single histidine residue from each monomer. The geometric coordination difference between CMP structures was evaluated by comparing the RMSD alignment over three atoms within the imidazole ring—C $\beta$ , C $\gamma$ , N $\delta$ 1, and N $\epsilon$ 2. Each CMP was then matched into all scaffolds of the subset library, as described above.

## 3.7 References

1. Lu Y, Yeung N, Sieracki N, Marshall NM: **Design of functional metalloproteins.** *Nature* 2009, **460**:855–862.
2. Waldron KJ, Rutherford JC, Ford D, Robinson NJ: **Metalloproteins and metal sensing.** *Nature* 2009, **460**:823–830.
3. Cheung WY, Regulators C: **Calmodulin plays a pivotal role in cellular regulation.** *Science* 1978, **1960**:19–27.

4. Umena Y, Kawakami K, Shen J-R, Kamiya N: **Crystal structure of oxygen-evolving photosystem II at a resolution of 1.9 Å.** *Nature* 2011, **473**:55–60.
5. Fasan R: **Tuning P450 enzymes as oxidation catalysts.** *ACS Catal* 2012, **2**:647–666.
6. Song WJ, Sontz PA, Ambroggio XI, Tezcan FA: **Metals in protein–protein interfaces.** *Annu Rev Biophys* 2014, **43**:409–431.
7. Karlin S, Zhu ZY, Karlin KD: **The extended environment of mononuclear metal centers in protein structures.** *Proc Natl Acad Sci USA* 1997, **94**:14225–14230.
8. Goodsell DS, Olson AJ: **Structural symmetry and protein function.** *Annu Rev Biophys* 2000, **29**:105–153.
9. Suzuki Y, Cardone G, Restrepo D, Zavattieri PD, Baker TS, Tezcan FA: **Self-assembly of coherently dynamic, auxetic, two-dimensional protein crystals.** *Nature* 2016, **533**:369–373.
10. Laganowsky A, Zhao M, Soriaga AB, Sawaya MR, Cascio D, Yeates TO: **An approach to crystallizing proteins by metal-mediated synthetic symmetrization.** *Protein Sci* 2011, **20**:1876–1890.
11. Weeratunga SK, Lovell S, Yao H, Battaile KP, Fischer CJ, Gee CE, Rivera M: **Structural studies of bacterioferritin B from *Pseudomonas aeruginosa* suggest a gating mechanism for iron uptake via the ferroxidase center.** *Biochemistry* 2010, **49**:1160–1175.

12. Smith MC, Mclendon G: **Functional design of the heme proteins: reactions with linear ligands.** *J Am Chem Soc* 1980, **4749**:5666–5670.
13. Desjarlais JR, Clarke ND: **Computer search algorithms in protein modification and design.** *Curr Opin Struct Biol* 1998, **8**:471–475.
14. Hellinga HW, Caradonna JP, Richards FM: **Construction of new ligand binding sites in proteins of known structure: II. Grafting of a buried transition metal binding site into Escherichia coli thioredoxin.** *J Mol Biol* 1991, **222**:787–803.
15. Shete VS, Benson DE: **Protein design provides lead (II) ion biosensors for imaging molecular fluxes around red blood cells.** *Biochemistry* 2009, **48**:462–470.
16. Yang W, Jones LM, Isley L, Ye Y, Lee HW, Wilkins A, Liu ZR, Hellinga HW, Malchow R, Ghazi M, Yang JJ: **Rational design of a calcium-binding protein.** *J Am Chem Soc* 2003, **125**:6165–6171.
17. Benson DE, Wisz MS, Liu W, Hellinga HW: **Construction of a novel redox protein by rational design: conversion of a disulfide bridge into a mononuclear iron-sulfur center.** *Biochemistry* 1998, **37**:7070–7076.
18. Roy A, Sarrou I, Vaughn MD, Astashkin AV, Ghirlanda G: *De novo design of an artificial bis[4Fe-4S] binding protein.* *Biochemistry* 2013, **52**:7586–7594.
19. Roy A, Sommer DJ, Schmitz RA, Brown CL, Gust D, Astashkin A, Ghirlanda G: **A de novo designed 2[4Fe-4S] ferredoxin mimic mediates electron transfer.** *J Am Chem Soc* 2014, **136**:17343–17344.

20. Scott MP, Biggins J: **Introduction of a [4Fe-4S (S-cys)<sub>4</sub>]+1,+2 iron-sulfur center into a four-alpha helix protein using design parameters from the domain of the Fx cluster in the photosystem I reaction center.** *Protein Sci* 1997, **6**:340–6.
21. Robertson DE, Farid RS, Moser CC, Urbauer JL, Mulholland SE, Pidikiti R, Lear JD, Wand AJ, DeGrado WF, Dutton PL: **Design and synthesis of multi-haem proteins.** *Nature* 1994, **368**:425–432.
22. Zaytsev DV, Morozov VA, Fan J, Zhu X, Mukherjee M, Ni S, Kennedy MA, Ogawa MY: **Metal-binding properties and structural characterization of a self-assembled coiled coil: formation of a polynuclear Cd-thiolate cluster.** *J Inorg Biochem* 2013, **119**:1–9.
23. Gibney BR, Isogai Y, Rabanal F, Reddy KS, Grosset AM, Moser CC, Dutton PL: **Self-assembly of heme A and heme B in a designed four-helix bundle: Implications for a cytochrome c oxidase maquette.** *Biochemistry* 2000, **39**:11041–11049.
24. Regan L, Clarke ND: **A tetrahedral Zinc(II)-binding site introduced into a designed protein.** *Biochemistry* 1990, **29**:10878–10883.
25. Grzyb J, Xu F, Weiner L, Reijerse EJ, Lubitz W, Nanda V, Noy D: **De novo design of a non-natural fold for an iron-sulfur protein: alpha-helical coiled-coil with a four-iron four-sulfur cluster binding site in its central core.** *Biochem Biophys Acta* 2010, **1797**:406–413.

26. Klemba M, Regan L: **Characterization of metal binding by a designed protein: single ligand substitutions at a tetrahedral Cys2His2 site.** *Biochemistry* 1995, **34**:10094–10100.
27. Müller HN, Skerra A: **Grafting of a high-affinity Zn(II)-binding site on the beta-barrel of retinol-binding protein results in enhanced folding stability and enables simplified purification.** *Biochemistry* 1994, **33**:14126–14135.
28. Eskandari V, Yakhchali B, Sadeghi M, Karkhane AA: **In silico design and construction of metal-binding hybrid proteins for specific removal of cadmium based on CS3 pili display on the surface of Escherichia coli.** *Biotechnol Appl Biochem* 2013, **60**:564–572.
29. Zastrow ML, Peacock AFA, Stuckey JA, Pecoraro VL: **Hydrolytic catalysis and structural stabilization in a designed metalloprotein.** *Nat Chem* 2011, **4**:118–123.
30. Plegaria JS, Dzul SP, Zuiderweg ERP, Stemmler TL, Pecoraro VL: **Apoprotein structure and metal binding characterization of a *de novo* designed peptide,  $\alpha$ 3DIV, that sequesters toxic heavy metals.** *Biochemistry* 2015, **54**:2858–2873.
31. Klemba M, Gardner KH, Marino S, Clarke ND, Regan L: **Novel metal-binding proteins by design.** *Nat Struct Biol* 1995, **2**:368–373.
32. Marino SF, Regan L: **Secondary ligands enhance affinity at a designed metal-binding site.** *Chem Biol* 1999, **6**:649–655.

33. Salgado EN, Radford RJ, Tezcan FA: **Metal-directed protein self-assembly**. *Acc Chem Res* 2010, **43**:661–672.
34. Zhou L, Bosscher M, Zhang C, Ozçubukçu S, Zhang L, Zhang W, Li CJ, Liu J, Jensen MP, Lai L, He C: **A protein engineered to bind uranyl selectively and with femtomolar affinity**. *Nat Chem* 2014, **6**:236–41.
35. Leaver-fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YA, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popovi Z, Havranek JJ, Karanicolar J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P: **Rosetta 3: an object-oriented software suite for the simulation and design of macromolecules**. *Methods Enzym* 2014, **487**:545–574.
36. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA, Röthlisberger D, Baker D: **New algorithms and an in silico benchmark for computational enzyme design**. *Protein Sci* 2006, **15**:2785–2794.
37. Mills JH, Khare SD, Bolduc JM, Forouhar F, Mulligan VK, Lew S, Seetharaman J, Tong L, Stoddard BL, Baker D: **Computational design of an unnatural amino acid dependent metalloprotein with atomic level accuracy computational design of an unnatural amino acid dependent metalloprotein with atomic level accuracy**. *J Am Chem Soc* 2013, **135**:13393–13399.
38. Guffy SL, Der BS, Kuhlman B: **Probing the minimal determinants of zinc binding with computational protein design**. *Protein Eng Des Sel* 2016, **29**:327–338.

39. Yeung N, Lin Y, Gao Y, Zhao X, Russell BS, Lei L, Miner KD, Robinson H, Lu Y: **Rational design of a structural and functional nitric oxide reductase** 2009, **462**:1079–1084.
40. Mills JH, Sheffler W, Ener ME, Almhjell PJ, Oberdorfer G, Perieira JH, Parmeggiani F, Sankaran B, Zwart PH, Baker D: **Computational design of a homotrimeric metalloprotein with a trisbipyridyl core.** *Proc Natl Acad Sci* 2016, **113**:15012–15017.
41. Tantillo DJ, Chen J, Houk KN: **Theozymes and compuzymes: theoretical models for biological catalysis.** *Curr Opin Chem Biol* 1998, **2**:743–750.
42. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The protein data bank.** *Nucleic Acids Res* 2000, **28**:235–242.
43. Meiler J, Baker D: **ROSETTALIGAND: protein – small molecule docking with full side-chain flexibility.** *Proteins Struct Funct Bioinforma* 2006, **548**:538–548.
44. Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D: **De novo enzyme design using Rosetta3.** *PLoS One* 2011, **6**:1–12.
45. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D: **Protein – protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations.** *J Mol Biol* 2003, **2836**:281–299.

46. Khatib F, Cooper S, Tyka MD, Xu K, Makedon I, Popovic Z, Baker D, Players F: **From the cover: Algorithm discovery by protein folding game players.** *Proc Natl Acad Sci* 2011, **108**:18949–18953.
47. DiMaio F, Leaver-Fay A, Bradley P, Baker D, André I: **Modeling symmetric macromolecular structures in Rosetta3.** *PLoS One* 2011, **6**:1–13.
48. Fleishman SJ, Baker D: **Perspective role of the biomolecular energy gap in protein design, structure, and evolution.** *Cell* 2012, **149**:262–273.
49. Hsin K, Sheng Y, Harding MM, Taylor P, Walkinshaw MD: **MESPEUS: a database of the geometry of metal sites in proteins.** *J Appl Crystallogr* 2008, **41**:963–968.
50. Hansen WA, Mills JH, Khare SD: **Computational design of multinuclear Metalloproteins using unnatural Amino Acids In: Methods in Molecular Biology.** Vol. 1414 *Humana Press Inc* 2016; pp. 173–185.
51. Chaudhury S, Lyskov S, Gray JJ: **PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta.** *Struct Bioinforma* 2010, **26**:689–691.

## 3.8 Supplementary Material

### 3.8.1 Symmetric Protein Recursive Ion Sampler (SyPRIS) package

*Step 1) Read a given backbone dependent rotamer library:* SyPRIS generates a backbone-independent Dunbrack rotamer library 52 from the backbone-dependent library in the Rosetta database. All torsion angles and standard deviations of each rotamer with the same identity across all backbone variations are averaged to make a single consensus rotamer with a reference identity key, which is stored in a dictionary. This procedure creates a backbone independent library with keys that correspond to identities obtained from the backbone dependent library.

*Step 2) Construction of the CMP object:* When a CMP is read into SyPRIS, all individual amino acids within the CMP are identified; the atom types and Cartesian coordinates of each amino acid are grouped together and labeled as a CMP branch. An atom connectivity map, similar to an atom tree in Rosetta 35, is constructed for each CMP branch—detailing the  $\chi$  torsion angles. The CMP branches are further separated into two parts whose torsion angles are either continuously or discretely sampled (Figure 5A). When sampling, for each rotamer key in the rotamer library, the torsion angles of a base rotamer are applied to the CMP branches. Additional "extra" rotamers can be generated by rotation about the relevant bond to adopt other values within the standard deviation of a given rotameric energy well determined from the rotamer library 52. Discretely sampled  $\chi$  torsions are restricted to torsion angle values obtained from these (base and extra) rotamers. Continuously sampled  $\chi$  torsion angles can adopt any value within the standard deviation window around a base rotamer. The atom that separates the CMP

branch into its discrete and continuously sampled parts is called the fulcrum atom and defaults to the  $\beta$  carbon atom type (Figure 5A). For example, if the CMP is comprised of two symmetrically coordinating glutamate residues and a metal center, each glutamate will become a CMP branch, each with its own atom connectivity map. If the beta carbon is the fulcrum atom, then the discretely sampled  $\chi$  torsions will be  $\chi_1$  and  $\chi_2$  and the continuously sampled  $\chi$  torsion is  $\chi_1$  (Figure 5A). With the atom map in place, all branch torsion angles are initially set to zero.

*Step 3) Generation of CMP base rotamer ensemble:* The CMP is translated to the geometric center of the scaffold and rotated such that compatible axes of symmetry are aligned to the scaffold axes of symmetry. Using a rotamer dictionary as described in Step 1, all base rotamer  $\chi$  torsion angles are applied inversely and stored within the SyPRIS object—generating an ensemble of backbone constellations compatible with metal chelation. The rotamer constellations are referred to as the ensemble of base rotamers and are labeled by the residue type and rotamer identifier. The number of discretely sampled inverse rotamers for any chosen sidechain torsion can be increased by sampling farther away from the base rotamer (increasing the overall "standard deviation" within which rotamer samples are chosen) and by sampling more finely in a given range of torsion angle values for any chosen torsion angle. This is controlled, in a manner analogous to RosettaMatch 46, by setting standard-deviation multiplier and periodicity parameters. For example, if the  $\chi_2$  for a histidine is  $90^\circ$  and the standard deviation obtained from the rotamer library is  $\pm 15^\circ$ , when the multiplier and periodicity are set to 2, SyPRIS generates 4 additional rotamers ( $\chi_2 = 60^\circ, 75^\circ, 105^\circ, \text{ or } 120^\circ$ ) in addition to the primary rotamer ( $\chi_2 = 90^\circ$ ) for a total of 5 base rotamers.

*Step 4) Locate potential residue match positions along the axis of symmetry:* For every base rotamer generated in Step 3, SyPRIS defines the CMP center as the geometric center between all atoms labeled as the fulcrum atom ( $C\beta$ ). The radial distance from the center to a fulcrum atom is used to determine potential matching residues within the scaffold. Scaffold atoms with the same atom type as the fulcrum atom are grouped by their sequence index—placing symmetric atoms from multiple chains within the same group. For each symmetric group of scaffold fulcrum atoms, the radial distance from the axis of symmetry is calculated and compared to the CMP radial distances of each rotamer ensemble member. For example, when the fulcrum atom is  $C\beta$ , if the radial distance for one CMP base rotamer is  $3.0\text{\AA}$ , then SyPRIS will locate all scaffold residues (grouped symmetrically) whose  $C\beta$  are within a  $3\pm 1\text{\AA}$  distance from the axis of symmetry. In C-symmetric sampling, the CMP center of each base rotamer is used to translate the CMP along the axis of symmetry to the centers of the symmetrically grouped scaffold residues within the accepted radial distance range. The CMP is then rotated about the axis of symmetry until one CMP fulcrum atom and the scaffold main chain fulcrum atom are at a minimum distance apart (Figure 1C).

*Step 5) Continuous  $\chi$  torsion sampling:* The backbone atom positions of the CMP and scaffold are aligned by continuously sampling the  $\chi$  torsion angles ( $\chi_1$  if the fulcrum atom is  $C\beta$ ) stemming from the fulcrum atom to the backbone nitrogen atom. The atom connectivity map generated in Step 2 is used to calculate an idealized torsion angle between the CMP and the scaffold atom positions. The ideal torsion angles are measured using two atoms from the CMP (including the fulcrum atom) and two atoms from the scaffold residue. For example if the continuously sampled torsion angle is  $\chi_1$  (fulcrum is

C $\beta$ ), then the idealized torsion angle would be the angle measured from the CMP C $\gamma$ -C $\beta$  and scaffold C $\alpha$ -N atoms.

The difference between the idealized and CMP branch torsion angle is applied to the CMP branch by rotating the CMP backbone atoms about a continuously sampled  $\chi$  torsion bond. The exact torsion angle difference is applied if the difference is within the standard deviation window of the rotamer ensemble member; otherwise, the torsion angle difference will be reduced to the maximum standard deviation value. Increasing the standard deviation multiplier will increase the maximum degree of rotation allowed. If the chelating residue has less than the number of atoms necessary to make an idealized torsion angle, for example serine and threonine, first the metal ion, and then an arbitrary point in space are used as surrogates. This process continues until all continuously sampled torsions have been adjusted—creating CMP matches. At the end of sampling, each CMP match is subject to a round of scaffold backbone atom clash checking. An atom-atom clash distance cutoff of 3.0Å was used in this study. CMP matches found to clash with the backbone atoms were discarded.

*Step 6) Scoring match outputs:* To score a placed CMP, we generated a term that considers the proximity and the directionality of the aligned CMP to scaffold backbone atoms. To determine the proximity of a CMP and scaffold backbone alignment we calculate the RMSD across all sampled atoms—C $\beta$ , C $\alpha$ , and N. To determine the directionality of the connected atom vectors we developed a metric that we call root-mean squared angle (RMSA) See (1) above. The RMSA value is summed over all backbone atom bonds. We chose the functional form in Eq. 1 for this term as it scales

linearly with RMSD from 0 to 1 and becomes less correlated as each term approaches 1 (Figure 5B). Intuitively, RMSA scales with a ratio equivalent to  $25^\circ$  and therefore a score of 1 implies the inter-bond angle deviates on average  $25^\circ$  across all atom-atom bonds while a score of 0.5 is equivalent to  $12.5^\circ$  average deviation. The SyPRIS score is the average of RMSA and RMSD summed over all peptide backbone bonds and atomic positions, respectively.

### 3.8.2 Run command

```
~/Rosetta/main/source/bin/rosetta_scripts.linuxgccrelease
@general.flags @my_flags.flags
```

### 3.8.3 Flags files

```
-database ~/Rosetta/main/database/-parser:protocol my_xml.xml-ex1-ex2-
score:weights talaris2013_cst-extrachi_cutoff 0-use_input_sc-
output_virtual-ignore_zero_occupancy false-overwrite
-ignore_unrecognized_res-suffix "_postmin"-s my_pdb_file.pdb-
parser:script_vars residue=## loop_range=(##-2)-(##+2)
sym_file=my_symm_file.symm cst_file=my_cst_file.cst
```

### 3.8.4 XML File

```
<ROSETTASCRIPTS>
<SCOREFXNS>
  <scorefxn1 weights=talaris2013_cst symmetric=1>
    <Reweight scoretype=coordinate_constraint weight=1.0
  /></scorefxn1>
  <scorefxn2 weights=enxdes_polyA_min.wts symmetric=1>
    <Reweight scoretype=coordinate_constraint weight=1.0
  /></scorefxn2>
</SCOREFXNS>
<RESIDUE_SELECTORS>
  <Index name=chelant_res resnums=%residue% />
  <Not name=not_chelant_res selector=chelant_res />
</RESIDUE_SELECTORS>
<TASKOPERATIONS>
  <InitializeFromCommandline name=init/>
  <IncludeCurrent name=keep_curr/>
  <DesignAround name=aroundCMP design_shell=4.0 resnums=%residue%
  repack_shell=8.0 allow_design=0 resnums_allow_design=0/> #set design
  true
  <OperateOnResidueSubset name=chelant selector= not_chelant_res>
    <PreventRepackingRLT/>
  </OperateOnResidueSubset>
```

```

</TASKOPERATIONS>
<FILTERS>
    <PoseInfo name=p_info />
    <TaskAwareScoreType name=atom_coord_res task_operations=chelant
scorefxn=scorefxn1 score_type=coordinate_constraint threshold=10000
mode=total />
    <TaskAwareScoreType name=total_res task_operations=chelant
scorefxn=scorefxn1 score_type=total_score threshold=10000 mode=total />
</FILTERS>
<MOVERS>
    <SetupForSymmetryname=setup_symm
definition="%%sym_file%%"/><FavorNativeResidue name=native bonus=1.5 />
    <ConstraintSetMover name=cstADD add_constraints=true
cst_file="%%cst_file%%" />
    <SymPackRotamersMover name=repack scorefxn=scorefxn1
task_operations=init,keep_curr,aroundCMP/><SymMinMover name=min
scorefxn=scorefxn1 type=lbfgs_armijo_nonmonotone tolerance=0.001 bb=0
chi=1 jump=0/>
    <SymMinMover name=firstMin scorefxn=scorefxn2
type=lbfgs_armijo_nonmonotone tolerance=0.001 bb=0 chi=1 jump=0/>
    <ParsedProtocol name=repack_minimize>
        <Add mover=repack/>
        <Add mover=min/>
    </ParsedProtocol>
</MOVERS>
<PROTOCOLS>
    <Add mover=setup_symm/>
    <Add mover=native/> #add for design
    <Add mover=cstADD/>
    <Add mover=firstMin/>
    <Add mover=repack_minimize/>
    <Add filter=total_res/>
    <Add filter=atom_coord_res/>
</PROTOCOLS>
</ROSETTASCRIPTS>

```

### 3.8.5 Path to referenced SyPRIS python scripts, wrappers, and preparatory scripts can

be obtained from RosettaCommons GitHub Repository:

~/Rosetta/main/source/src/apps/pilot/wah49/

### 3.8.6 Supplementary Tables

Table 1-3. Table S1.

The benchmark PDB codes, metal identifications, and corresponding chelating residues acquired from the Mespeus database.

	PDB	Metal type	Metal ID	Residue Type	Residue Index	Oligomeric Chains
1	1AOR	FE	606A	GLU, HIS	332, 383	A, B
2	1BCH	NA	5B	HIS	99	A(1), B(2), C(3)
3	1C3H	CA	8001D, 8002E, 8003F	ASP	198	D, E
4	1CQX	NA	490A	GLU	100	A, B
5	1E2A	MG	106C	ASP	81	A, B, C
6	1EF8	NI	300A	HIS	220	A, B, C
7	1EM9	MG	306A	ASP	71	A, B
8	1HFE	ZN	500S	ASP	86	S, T
9	1HWT	ZN	138D	HIS	80, 91	D, H
10	1KXG	MG	701A	GLN	234	A, B, C
11	1NC7	MG	1504A	THR	104	A, B, C, D
12	1NS7	NA	1401A	ASP	13	A, B
13	1PYZ	CO	101A	HIS	5	A, B
14	1Q5H	MG	999A	GLU	112	A, B, C
15	1Q7B	CA	9002A	GLU	233	A, D
16	1RXC	K	2101A	GLU, SER	49, 73	A, B
17	1SAW	MG	226A	ASN	19	A, B
18	1T92	ZN	141A	GLU	44	A, B
19	1TXK	NA	600A	SER	320	A, B
20	1ULI	FE	700B	HIS	24	B, D, F
21	1VRG	MG	516A	HIS	379	A, D
22	1VZH	CA	1128A	THR	90	A, B
23	1WAA_1	ZN	1090A	HIS	20	A, E
24	1WAA_2	ZN	1092A	ASP	29	A, F.B
25	1Y7W	ZN	285B	HIS	13	A, B
26	1Y9D	NA	2605A	GLN	455	A, C
27	1YU4	MG	2001A, 2002A	GLU	312	A, B, C
28	1Z6O	CA	5302A	GLN	161, 164	A, E, L
29	1ZDN	NA	157A	ASN	124	A, B
30	1ZEI	ZN	54C	HIS	10	A, C, E
31	2A2O	K	248B	GLU	128	A, B
32	2C5U	CA	1378B, 1379B	ASP	212	A, B

Table 2-3. S1. Continued.

	<b>PDB</b>	<b>Metal type</b>	<b>Metal ID</b>	<b>Residue Type</b>	<b>Residue Index</b>	<b>Oligomeric Chains</b>
<b>33</b>	2CG4	MG	1154A	ASP	58	A, B
<b>34</b>	2CKI	CA	1403A	ASP	152	A, B
<b>35</b>	2DGE	ZN	1001A	ASP, GLU	74, 159	A, C
<b>36</b>	2ERV	CA	195A	ASP	106	A, B
<b>37</b>	2ESL	CA, ZN	1A, 4A	ASP, HIS	93, 174	A, B, C
<b>38</b>	2FCA	K	252A	SER	167	A, B
<b>39</b>	2FV2	MN	900A	GLN, ASP	136, 176	A, C
<b>40</b>	2GVH	NA	277B	ASP	157	A, B, C
<b>41</b>	2HZY	CA	1206A	ASP	123	A, B
<b>42</b>	2I2X	ZN	524A	HIS, GLU	318, 320	A, C
<b>43</b>	2J4J	CO	230A	HIS	104	A, C, E
<b>44</b>	2MPR	CA	431A	ASP	78	A, B, C
<b>45</b>	2OGA	NA	1022A	ASN	65	A, B
<b>46</b>	2P2L	ZN	901A	ASP	38	A, B, C
<b>47</b>	2PC5	MG	167A	GLU	138	A, B, C
<b>48</b>	2UVE	NI	1607B	HIS	97	A, B
<b>49</b>	2V21	NA	1069A	GLU	19	A, B, C
<b>50</b>	2VBL	MG	1026C	ASP	20	A, B
<b>51</b>	2VQG	ZN	1092C	GLU	37	C, H
<b>52</b>	2VRS	ZN	1328A	HIS	158	A, B, C
<b>53</b>	2W37	NI	1344A	HIS	79	A, B, C
<b>54</b>	2WD6	K	1762B	THR	712	A, B
<b>55</b>	2WMY	NI	1149A	HIS	47	A, D
<b>56</b>	3BF3	MG	247A	HIS	42	A, D, F
<b>57</b>	3BIX	NI	2A	HIS	640, 642	A, D
<b>58</b>	3CVJ	MG	243A	HIS, GLU	53, 57	A, B
<b>59</b>	3DE8	CA	109A	ASP	74	A, C
<b>60</b>	3F5J	ZN	1001A	ASP, HIS	65, 68	A, B
<b>61</b>	3FSX	MG	332A	GLU	183	A, B, C
<b>62</b>	3GSD	NA	120E, 121B	GLN	32	B, E
<b>63</b>	3GTD	NA	463A	ASN	343	A, B
<b>64</b>	3GXZ	MG	102A	GLN	50	A, B
<b>65</b>	3ICF	NA	515A	SER, THR	219, 223	A, B
<b>66</b>	3K9U	NI	161A	ASN	100	A, B
<b>67</b>	3KC2	MG	356B	ASN	139	A, B

Table 3-3. Table S2.

The RMSD results for the perfectly symmetric Rosetta-generated model aligned to the raw PDB structure.

	PDB	Chelating Residue(s)	Residue RMSD	Scaffold RMSD		PDB	Chelating Residue(s)	Redidue RMSD	Scaffold RMSD
1	1AOR	332, 383	0.8	0.2	43	2AQT	73, 133	2.9	0.5
2	1ASO	286	3.8	0.3	44	2C36	39, 215	1.2	0.3
3	1BCH	99	1.5	0.5	45	2C5U	212	0.1	0.4
4	1C3H	198	0.9	0.5	46	2CG4	58	0.3	0.1
5	1CQX	100	0.6	0.6	47	2CKI	152	0.2	0.1
6	1E2A	81	0.9	0.5	48	2DGE	74, 159	0.6	0.6
7	1EF8	220	1.5	0.7	49	2EJN	89	3.0	0.7
8	1EM9	71	0.6	0.5	50	2ERV	106	0.2	1.0
9	1GMW	96	2.7	2.2	51	2ESL	93, 174	1.8	0.6
10	1HFE	86	0.3	0.3	52	2EUL	45	3.6	0.3
11	1HWT	80, 91	0.5	0.2	53	2FCA	167	0.1	0.7
12	1KXG	234	0.9	0.6	54	2FV2	136, 176	1.2	1.1
13	1MFT	41	1.2	0.9	55	2GVH	157	0.3	0.6
14	1NC7	104	0.6	0.5	56	2HZY	123	0.1	0.8
15	1NS7	13	0.2	0.5	57	2I2X	318, 320	0.9	0.2
16	1PYZ	5	1.1	0.5	58	2J4J	104	0.3	0.7
17	1Q5H	112	0.9	1.2	59	2MPR	78	0.3	0.1
18	1Q7B	233	0.1	0.4	60	2OGA	65	0.4	0.5
19	1RHF	166	2.4	0.8	61	2P2L	38	1.0	1.1
20	1RXC	49, 73	0.9	1.4	62	2PC5	138	1.0	0.5
21	1S4I	71, 137	2.8	1.0	63	2UVE	97	1.0	0.8
22	1S6B	112	1.3	1.3	64	2V21	19	0.3	0.2
23	1SAW	19	0.3	0.4	65	2VBL	20	0.2	0.5
24	1T92	44	0.4	0.3	66	2VQG	37	0.4	0.5
25	1TU4	83	25.3	7.1	67	2VRS	158	0.8	0.6
26	1TXK	320	1.1	0.5	68	2W37	79	0.4	0.2
27	1U10	147, 150	1.1	1.1	69	2WD6	712	0.6	0.4
28	1ULI	24	0.6	0.2	70	2WMY	47	0.8	0.8
29	1V9B	15	3.2	0.6	71	3BF3	42	0.3	0.4
30	1VRG	379	0.4	0.2	72	3BIX	640, 642	0.3	0.0
31	1VZH	90	0.1	0.3	73	3CVJ	53, 57	0.8	0.8
32	1WAA_1	29	0.6	0.7	74	3DE8	74	0.2	0.6
33	1WAA_2	20	0.6	0.5	75	3F5J	65, 68	0.2	0.1
34	1XU1	106	1.6	0.3	76	3FSX	183	0.9	0.6
35	1Y6P	71, 92	3.6	2.5	77	3GHZ	149	1.4	0.4
36	1Y7W	13	0.4	0.5	78	3GSD	32	0.4	0.7
37	1Y9D	455	0.3	0.3	79	3GTD	343	0.3	0.3
38	1YU4	312	0.4	0.1	80	3GXZ	50	1.3	0.5

Table 4-3. S2. Continued.

	PDB	Chelating Residue(s)	Residue RMSD	Scaffold RMSD		PDB	Chelating Residue(s)	Redidue RMSD	Scaffold RMSD
39	1Z6O	161, 164	0.9	0.1	81	3ICF	219, 223	0.2	0.0
40	1ZDN	124	1.1	1.1	82	3IKP	232	1.9	1.2
41	1ZEI	10	1.6	1.1	83	3K9U	100	0.8	0.8
42	2A2O	128	0.7	0.3	84	3KC2	139	0.5	0.9

Value outside the threshold

Excluded from the benchmark

## Chapter 4:      Enzyme stabilization via computationally guided protein stapling

### 4.1 Preface

A version of this chapter has been published in *Proceedings of the National Academy of Sciences USA* and is formatted in the journal style.

### 4.2 Abstract

Thermostabilization represents a critical and often obligatory step toward enhancing the robustness of enzymes for organic synthesis and other applications. While directed evolution methods have provided valuable tools for this purpose, these protocols are laborious and time-consuming and typically require the accumulation of several mutations, potentially at the expense of catalytic function. Here, we report a minimally invasive strategy for enzyme stabilization that relies on the installation of genetically encoded, nonreducible covalent staples in a target protein scaffold using computational design. This methodology enables the rapid development of myoglobin-based cyclopropanation biocatalysts featuring dramatically enhanced thermostability ( $\Delta T_m = +18.0\text{ }^{\circ}\text{C}$  and  $\Delta T_{50} = +16.0\text{ }^{\circ}\text{C}$ ) as well as increased stability against chemical denaturation [ $\Delta C_m$  (GndHCl) = 0.53 M], without altering their catalytic efficiency and stereoselectivity properties. In addition, the stabilized variants offer superior performance and selectivity compared with the parent enzyme in the presence of a high concentration of organic cosolvents, enabling the more efficient cyclopropanation of a water-insoluble substrate. This work introduces and validates an approach for protein stabilization which should apply to a variety of other proteins and enzymes.

### 4.3 Introduction

Enzymes play a significant role in biotechnology and constitute attractive catalysts for the implementation of efficient, selective, and sustainable processes for the production of pharmaceuticals, fine chemicals, and biofuels.<sup>1,2</sup> The marginal stability of proteins, however, poses a fundamental challenge for the exploitation of enzymes for practical-scale syntheses and chemical manufacturing, which often require harsh reaction conditions such as elevated temperature and exposure to organic solvents.<sup>1</sup> Due to these limitations, protein stabilization against thermal and chemical denaturation has represented a long-standing goal in enzyme design and engineering. In addition to higher robustness to operational conditions, increasing the thermostability of a protein can enhance its evolvability (i.e., its tolerance to mutagenesis for the acquisition of new or improved functions).<sup>3</sup>

Well-known experimental strategies for enzyme stabilization include directed evolution and consensus mutagenesis.<sup>4-6</sup> While useful in many instances, these procedures remain laborious and time-consuming, typically requiring several rounds of mutagenesis and screening to obtain variants with significantly increased stability (i.e.,  $\Delta T_m > 5-10$  °C).<sup>7</sup> Notable examples of enzyme stabilization have been reported using computational design.<sup>8</sup> In this area, methods have focused on optimizing native state interactions, for example through improving core packing,<sup>9-11</sup> fragment contacts,<sup>12</sup> combined structure- and phylogeny-guided energy optimization,<sup>13,14</sup> surface charge optimization,<sup>15,16</sup> and rigidification.<sup>17, 18</sup> A complementary but comparatively less explored approach has focused on decreasing the configurational entropy of the unfolded state ensemble,<sup>19</sup>

primarily via the installation of nonnative disulfide linkages.<sup>20</sup> Indeed, the introduction of nonnative disulfide bridges<sup>20</sup> has proven useful for increasing the thermostability of enzymes, although stability–activity trade-offs were also observed.<sup>21, 22</sup> Disulfide cross-linking can additionally provide kinetic stabilization.<sup>23</sup> Unfortunately, the redox instability of disulfide bridges make this approach unsuitable for many proteins and enzymes that are meant to function in reducing environments, such as the intracellular milieu and/or in the presence of reductants.<sup>24–26</sup> Moreover, the reversibility and homotypic nature of disulfide bridges can complicate the folding of variants with multiple cross-links due to disulfide scrambling.<sup>27</sup> Thus, the development of computational strategies for protein stabilization by means of chemically stable covalent bonds remains an unmet challenge.

Here, we report the development of Rosetta-guided protein stapling (R-GPS), a method for enzyme/protein stabilization based on the computational design of genetically encodable, covalent "staples" in a protein of interest. This method utilizes the Rosetta enzyme design framework to identify optimal sites for intramolecular cross-linking of the protein scaffold via redox-stable and irreversible thioether bonds generated upon the reaction between a cysteine and a genetically encoded noncanonical amino acid containing a cysteine-reactive side-chain group (O-2-bromoethyl tyrosine, or O2beY).<sup>28</sup> The R-GPS approach was implemented and validated in the context of a myoglobin (Mb)-based cyclopropanation biocatalyst, resulting in the development of "cyclopropanases" with significantly increased stability against thermal and chemical denaturation as well as improved catalytic performance under harsh reaction conditions.

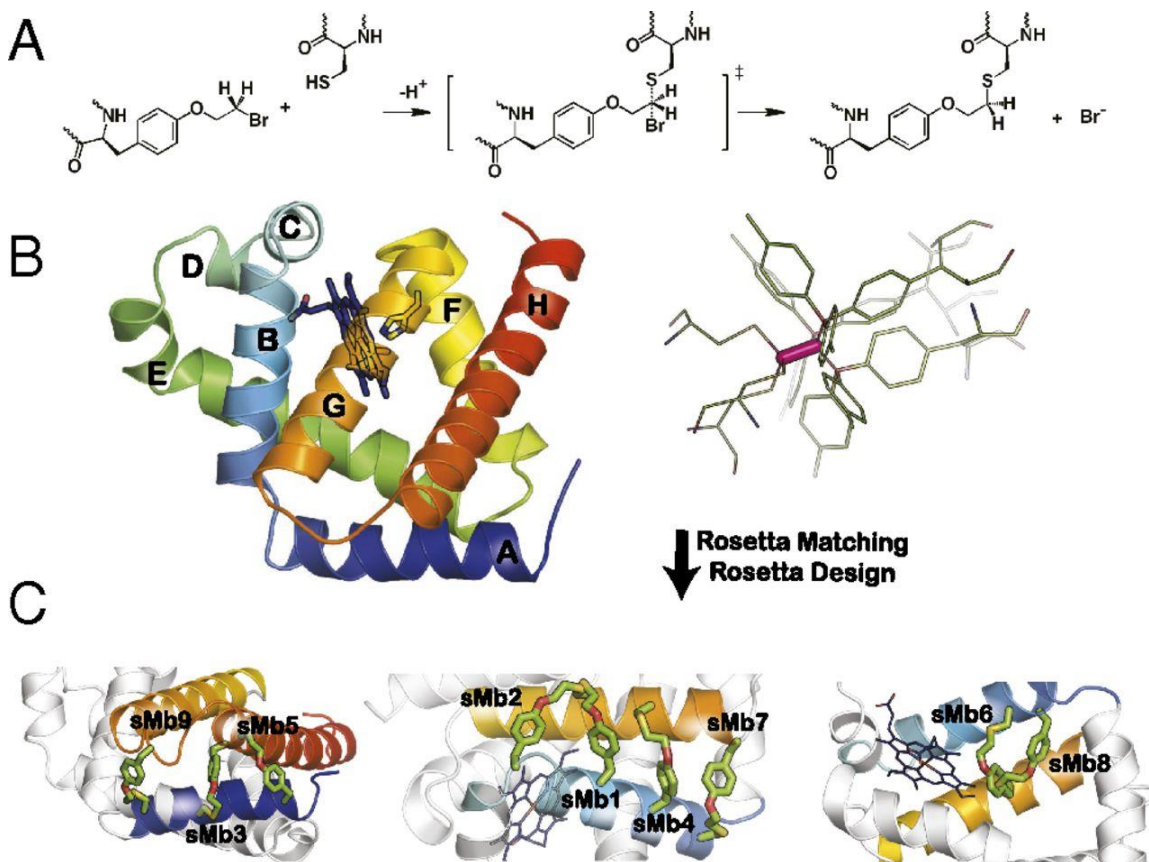
## 4.4 Results

### 4.4.1 Choice of Target Enzyme and Protein Stapling Chemistry.

A recently reported Mb-derived cyclopropanation biocatalyst, Mb(H64V,V68A),<sup>29</sup> were selected as the testbed for the development of R-GPS. Compared with wild-type Mb, Mb(H64V,V68A) bears two active site mutations that confer enhanced catalytic activity as well as high diastereo- and enantioselectivity (>90–99% *de* and *ee*) in the cyclopropanation of styrenes and vinyl arenes with ethyl  $\alpha$ -diazoacetate as the carbene donor.<sup>29</sup> On one hand, stabilization of Mb(H64V,V68A) was desirable for enhancing its robustness for synthetic applications, which include the synthesis of cyclopropane-containing drugs and other carbene transfer reactions.<sup>30, 31</sup> On the other hand, the high stereoselectivity of Mb(H64V,V68A) was envisioned to provide a sensitive probe for evaluating the impact of the computationally designed covalent staples on subtle functional properties such as stereoinduction. Indeed, previous studies indicated that the stereoselectivity of Mb-based cyclopropanation catalysts is highly sensitive to small structural variations in their active sites.<sup>30</sup>

The thioether bond-forming reaction between cysteine and the genetically encodable O-2-bromoethyl-tyrosine (O2beY)<sup>28, 32</sup> were selected as the protein cross-linking strategy (Fig. 1A). Based on these previous studies, the O2beY/Cys reaction was anticipated to offer several promising features for protein stapling, namely (i) the ability to mediate the spontaneous formation of the thioether cross-link at the posttranslational level, (ii) high chemoselectivity of O2beY toward cysteine-mediated alkylation in the presence of other nucleophilic residues (e.g., Lys and His), and (iii) "spatially controlled" reactivity,

whereby the O2beY and Cys undergo the nucleophilic substitution reaction only when located in close proximity.



**Figure 13-4.** Figure 1. Computational design approach.

(A) Stapling reaction between the noncanonical amino acid *O*-2-bromoethyl tyrosine and cysteine resulting in a chemically stable thioether bond. (B) Structure of Mb (Protein Data Bank ID code 1JP9) highlighting helices A–H (*Left*). The active-site heme group and metal-coordinating histidine residue are shown as sticks. The conformational ensemble of the modeled thioether linkage used to find compatible locations for stapling (*Right*). (C) Models showing locations of covalent staples between A–H and B–G helices in designs sMb1 through sMb9.

#### 4.4.2 Computational Design of Stapled Mb(H64V,V68A) Variants Using R-GPS.

As a first design principle we reasoned that cross-linking residues that are distal in the primary sequence but proximal in space in the folded state would maximize

thermostabilization by decreasing the chain entropic cost of folding.<sup>33-36</sup> As a second criterion we envisioned the need to identify compatible backbone positions for placing the haloalkane (O2beY) and thiol (Cys) side-chain groups so that the cross-link is accommodated in a strain-free configuration while making energetically favorable interactions with its surrounding residues.<sup>33</sup> Following these guiding principles, the N-terminal helices A and B and the C-terminal helices G and H of the target enzyme, Mb(H64V,V68A), were chosen as the target regions for the installation of the thioether bridges (Fig. 1B), as cross-links between these structural elements would bear the highest contact order. The RosettaMatch algorithm<sup>37</sup> was then applied to identify positions where the O2beY and Cys residues can be accommodated for the formation of the thioether staple. Once geometrically feasible backbone positions of the cross-linking residues were obtained (Fig. 1B), RosettaDesign<sup>38</sup> was performed to identify additional sequence changes to minimize steric clashes with the modeled staple. To assess the method, we selected for experimental characterization a diverse set of nine designs based on (i) the solvent accessibility of the staple in the protein structure, (ii) total number of amino acid substitutions, and (iii) Rosetta energies (Table 1). These designs were termed sMb1 through sMb9 and had differences in scores relative to the parent protein Mb(H64V,V68A) ranging from -6.7 (sMb1) to +16.6 (sMb9) Rosetta energy units (Reu), and intervening segment lengths ranging from 73 (sMb2) to 121 (sMb5) residues. Seven of designed variants feature a surface exposed thioether linkage, whereas for the remaining two (sMb6 and sMb8), the cross-link is buried (Fig. 1C). The number of designed amino acid substitutions, excluding the Cys and O2beY residues, ranged from zero (sMb2 and sMb5) to two (sMb3, sMb6, sMb7, sMb8, and sMb9).

Table 5-4. Computational and experimental values for Mb(H64V,V64A) and its variants

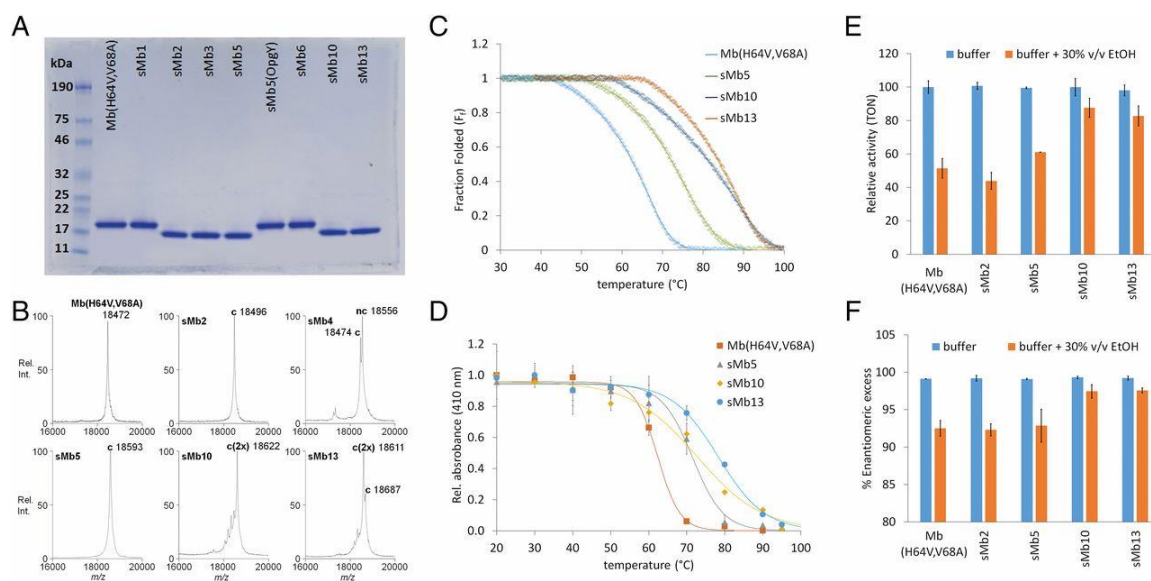
Variant	Mutations	$\Delta G^{\text{comp}}$ , Reu	Location staple <sup>a</sup>	Attack angle, deg <sup>b</sup>	Staple (MS)	$T_m$ °C ( $\Delta T_m$ ) <sup>c</sup>	$T_{50}$ , °C ( $\Delta T_{50}$ ) <sup>d</sup>	$C_m$ (M) ( $\Delta C_m$ ) <sup>e</sup>
Mb(H64V,V68A)	—	0	None	n.a.	n.a.	66.0 ± 1.0 (0.0)	62.3 ± 0.6 (0.0)	1.55 ± 0.02 (0.0)
sMb1	R31(O2beY), S35K, E109C	-6.73	B-G	102	No	67.5 ± 1.3 (+1.5)	59.1 ± 0.6 (-3.2)	n.d.
sMb2	H36(O2beY), E109C	-5.05	B/C-G	142	Yes	73.2 ± 0.2 (+7.2)	63.8 ± 0.3 (+1.5)	n.d.
sMb3	L9R, H12C, D122T, A127(O2beY)	-2.66	A-H	126	Yes	72.2 ± 1.4 (+6.2)	60 ± 3 (-2.3)	n.d.
sMb4	D27(O2beY), H113C, V114G	-0.76	B-G	98	Yes (partial)	56.1 ± 1.1 (-9.9)	50 ± 2 (-12.3)	n.d.
sMb5	G5(O2beY), D126C	-0.68	A-H	169	Yes	76.0 ± 2.0 (+10.0)	71.5 ± 0.9 (+9.2)	1.79 ± 0.02 (+0.24)
sMb6	G25C, I28A, L69A, I111(O2beY)	+1.70	B-G (core)	132	No	53.0 ± 1.0 (-13.0)	51.2 ± 0.3 (-11.1)	n.d.
sMb7	D20S, G23C, D27A, R118(O2beY)	+2.34	B-G	156	Yes	69.9 ± 1.4 (+3.9)	54 ± 1 (-8.3)	n.d.
sMb8	H24(O2beY), I28S, L69S, I111C	+3.05	B-G (core)	108	No	49.2 ± 1.0 (-16.8)	43.3 ± 0.6 (-19.0)	n.d.
sMb9	K16C, H119A, G121(O2beY), D122S	+16.6	A-G/H	112	No	64.3 ± 1.3 (-1.7)	52 ± 2 (-10.3)	n.d.
sMb10	G5(O2beY), H36(O2beY), E109C, D126C	n.a.	B/C-G + A-H	n.a.	Yes (double)	82.8 ± 0.8 (+16.8)	73 ± 2 (+10.7)	2.01 ± 0.05 (+0.46)
sMb11	G5(O2beY), H36(O2beY), F106A, E109C, D126C	n.a.	B/C-G + A-H	n.a.	Yes (double)	n.d.	61 ± 1 (-1.3)	n.d.
sMb12	G5(O2beY), H36(O2beY), E109C, D126C, G129E	n.a.	B/C-G + A-H	n.a.	n.d.	n.d.	66.6 ± 0.8 (+4.3)	n.d.
sMb13	G5(O2beY), H36(O2beY), E109C, H113E, D126C	n.a.	B/C-G + A-H	n.a.	Yes (double)	84.0 ± 0.2 (+18.0)	78.3 ± 0.8 (+16.0)	2.08 ± 0.03 (+0.53)

#### 4.4.3 Expression and Characterization of the sMb Designs.

All of the designed Mb(H64V,V68A)-derived variants could be expressed in soluble form from *Escherichia coli* cells containing an orthogonal aminoacyl-tRNA synthetase/tRNA<sub>CUA</sub> pair for O2beY incorporation<sup>28</sup> via amber stop codon suppression.<sup>39</sup> In addition, the constructs were all able to bind heme and fold properly, as evinced from the characteristic Soret band (~410 nm) in the corresponding UV-visible spectra (SI Appendix, Fig. S4).

SDS/PAGE analysis revealed an increase in electrophoretic mobility for five out of the nine protein constructs, namely sMb2, sMb3, sMb4, sMb5, and sMb7, compared with Mb(H64V,V68A) (Fig. 2A and SI Appendix, Fig. S5). This behavior is indicative of a more compact structure of the protein under denaturing and reducing conditions (SDS + DTT), which is consistent with the presence of the nonreducible thioether (O2beY/Cys) cross-link. These conclusions were corroborated by MALDI-TOF MS (Fig. 2B and SI Appendix, Fig. S6), which showed a single signal corresponding to the expected mass of these proteins minus 82 Da, deriving from the loss of HBr as a result of the O2beY/Cys cross-linking reaction. For sMb4, two  $m/z$  signals consistent with the stapled and unstapled form of the protein indicated the O2beY/cysteine cross-linking reaction had occurred only partially. For the remaining designs, the only observable species in the MS spectra corresponded to the O2beY-containing protein lacking the thioether cross-link. Notably, the integrity of the O2beY residue in these unstapled constructs demonstrates the lack of undesirable reactivity toward hydrolysis or abundant thiol-containing intracellular metabolites such as glutathione. These results also show how placing the

O2beY/Cys pair in close spatial proximity in the context of a folded protein is a necessary but not sufficient condition for productive stapling. The positional dependence of the stapling reaction could be rationalized based on the analysis of the rotamers of Cys and O2beY residues (Fig. 1B) in models of the sMb constructs in their unstapled forms. In particular, a good agreement was found between the experimental results and the accessibility of near-attack conformations (NACs) compatible with a bimolecular nucleophilic substitution reaction between the thiol group of the cysteine and the alkyl bromide group in O2beY (see SI Appendix for further discussion).



**Figure 14-4.** Figure 2. Characterization of stapled Mb(H64V,V68A) variants (sMb variants).

(A and B) SDS/PAGE gel (A) and MALDI-TOF MS spectrum (B) of Mb(H64V,V68A) and representative sMb variants. c, cross-linked; c(2×), doubly cross-linked; nc, not cross-linked. Calculated masses: Mb(H64V,V68A): 18,474 Da; sMb2 (c): 18,500 Da; sMb4 (c): 18,472 Da; sMb5 (c): 18,594 Da; sMb10 [c(2×)]: 18,621 Da; sMb13 [c(2×)]: 18,612 Da. (C) Thermal denaturation curves for Mb(H64V,V68A) and selected stapled variants as measured via CD at 220 nm ( $T_m$  determination). (D) Heat-induced inactivation curves (heme loss) for the same proteins as determined by the decrease of Soret band signal (408 nm) after incubation (10 min) at variable temperatures ( $T_{50}$  determination). See SI Appendix. Figs. S5-S7 for additional data. (E and F) Catalytic activity (E) and enantioselectivity (F) of Mb(H64V,V68A) and stapled

variants in styrene cyclopropanation reactions with EDA in buffer only and in the presence of 30% vol/vol ethanol. Relative activities refer to normalized catalytic turnovers (TON) relative to TON measured with Mb(H64V,V68A) in buffer only reactions. See SI Appendix, Figs. S10-S12 for related data with other organic cosolvents. Rel Int, relative intensity.

#### 4.4.4 Thermostability of the sMb Designs.

The thermostability of the Mb variants was examined by measuring their melting temperatures ( $T_m$ ) using CD (SI Appendix, Fig. S7). Mb(H64,V68A) was found to exhibit an apparent  $T_m$  of 66.0 °C (Fig. 2C), which is about 14 °C lower than that of wild-type Mb. Notably, all of the sMb constructs containing the thioether staple showed an increase in thermostability compared with the parent protein, with  $\Delta T_m$  values ranging from 3.9 °C to 10 °C (Table 1). The largest thermostabilization effect was observed for sMb5 ( $\Delta T_m$ : +10.0 °C), followed by sMb2 ( $\Delta T_m$ : +7.2 °C). In contrast, the sMb variants lacking the cross-link show comparable (sMb1 and sMb9) or lower  $T_m$  values (sMb4, sMb6, and sMb8) compared with Mb(H64V,V68A) ( $\Delta T_m$ : +1.5 to -16.8 °C; Table 1). To further examine the impact of the thioether staple, analogs of the two most stable sMb variants, sMb2 and sMb5, were prepared by substituting O2beY for O-propargyl-tyrosine (OpgY).<sup>40</sup> OpgY is an isostere of O2beY but is unable to react with cysteine to form the thioether cross-link, as confirmed by SDS/PAGE (Fig. 2A) and MALDI-TOF MS analyses (SI Appendix, Fig. S6). The resulting OpgY-containing variants, sMb2(OpgY) and sMb5(OpgY), displayed significantly lower  $T_m$  values than their stapled counterparts ( $\Delta T_m$ : -12.2 °C and -11.9 °C, respectively), confirming the stabilizing effect of the thioether cross-link, as encoded in computational design.

As a second measure of thermostability, half-maximal denaturation temperatures ( $T_{50}$ ) were determined by monitoring heme loss ( $\lambda_{\text{max}}$ : 408 nm) upon incubation of the hemoproteins (10 min) at variable temperatures (Fig. 2D). This is a more stringent assay of thermal stability since it monitors the ability of the sMb variants to remain associated with the heme cofactor, which is essential for their activity as cyclopropanation biocatalysts. As thermal denaturation of holomyoglobin is irreversible, an increase in  $T_{50}$  is also a measure of kinetic stabilization upon cross-linking. In this assay, Mb(H64V,V68A) showed a  $T_{50}$  value of 62.3 °C (Table 1). Whereas sMb3 and sMb7 showed reduced  $T_{50}$  values compared with Mb(H64V,V64A) (Table 1), both sMb5 and sMb2 exhibited improved thermal stability compared with the parent protein ( $\Delta T_{50}$ : +9.2 °C and +1.5 °C, respectively; Fig. 2C). These results demonstrate that the thermostabilization effect induced by the O2beY/Cys cross-links in both sMb2 and sMb5 is not restricted to the protein secondary structure, as determined by the CD melting curves, but extends to the heme-bound ("holo") forms of these proteins.

#### 4.4.5 Design and Characterization of Doubly Stapled Mb Variants.

Further stabilization of Mb(H64V,V68A) was then pursued by combining the covalent staples from the two most promising variants, namely sMb2 and sMb5, resulting in the sMb10 design. Upon production in *E. coli*, MS analysis confirmed the successful formation of both cross-links within this protein (Fig. 2B). Further characterization of sMb10 showed a nearly additive effect of the two thioether staples toward increasing thermostability of the hemoprotein in terms of both  $T_m$  and  $T_{50}$  ( $\Delta T_m$ : +16.8 °C;  $\Delta T_{50}$ : +10.7 °C; Table 1). Comparison of the CD spectra corresponding to sMb5, sMb10, and

Mb(H64V,V68A) revealed no major changes in the protein secondary structure as a result of the presence of a single (sMb5) and double staple (sMb10) (SI Appendix, Fig. S8).

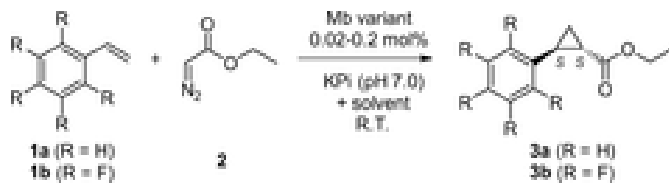
Evaluation and further stabilization of sMb10 were then pursued via structure-guided point mutagenesis. Investigation of the F106A substitution in sMb11 confirmed the energetic importance of the computed  $\pi$ - $\pi$  stacking interaction between the O2beY residue in position 36 and the phenyl ring of Phe106 (see SI Appendix for further discussion). Next, we noticed that the installation of the 36–109 cross-link (from sMb2) and 5–126 cross-link (from sMb5) replaces in each case a solvent-exposed negatively charged residue (i.e., Glu109 and Asp126, respectively) with the less-polar O2beY/Cys staple. Accordingly, two additional constructs (sMb12 and sMb13) were designed to incorporate "compensating" neutral-to-negatively charged amino acid substitutions (i.e., G129E and H113E, respectively) at a compatible position in proximity of each staple. Whereas sMb12 showed reduced thermostability ( $T_{50}$ ) compared with sMb10, sMb13 was found to exhibit higher  $T_m$  ( $\Delta T_m$ : +1.2 °C) as well as significantly increased stability against temperature-induced heme loss ( $\Delta T_{50}$ : +5.3 °C) compared with sMb10. The presence of the two O2beY/Cys cross-links in sMb13 was confirmed by MS analysis, which also revealed the presence of a small fraction of singly stapled protein (Fig. 2B). The CD spectra of sMb13 and sMb10 were found to be superimposable, indicating that the "charge-compensating" mutation H113E had minimal impact on the protein structure (SI Appendix, Fig. S8). Thus, as a result of merely five solvent-exposed mutations in sMb13, the thermostability ( $T_m$ ) of the original Mb(H64V,V64A) biocatalyst could be

augmented by +18 °C. These modifications also translated into a significantly increased thermal stability of the heme-bound form of the protein ( $\Delta T_{50}$ : +16 °C).

#### 4.4.6 Catalytic Activity and Selectivity of Stapled Mb(H64V,V68A) Variants in Cyclopropanation Reactions.

In the absence of selective pressure to maintain function, thermostabilization via protein engineering is often accompanied by a reduction in catalytic efficiency and/or selectivity of the target enzyme.<sup>41-43</sup> To assess the impact of the thioether staples on the catalytic properties and stereoselectivity of the thermostabilized sMb variants these biocatalysts were tested in a model cyclopropanation reaction with styrene (**1**) and  $\alpha$ -diazoacetate (EDA, **2**) (Table 2). Under the applied conditions [10 mM styrene, 20 mM EDA, 2  $\mu$ M Mb variant (0.02 mol%)] Mb(H64V,V68A) was found to produce (1*S*,2*S*)-ethyl 2-phenylcyclopropanecarboxylate (**3a**) in 94% yield (4,710 turnovers or TON) and high diastereomeric (99.4% *de*) and enantiomeric excess (99.1% *ee*) (entry 1, Table 2). Under identical conditions, both sMb10 and sMb13 were able to catalyze the formation of **3a** with equally high catalytic activity (92–94% yields; 4,710–4,615 TON) and stereoselectivity (>99% *de* and *ee*; entries 4–5, Table 2). These data indicated that thermostabilization induced by the stapling procedure had no negative impact on the carbene transfer reactivity of these hemoproteins, nor did it perturb the asymmetric environment provided by the distal heme pocket, which is critical for inducing the (1*S*,2*S*)-enantioselectivity during the cyclopropanation reaction.<sup>29</sup>

**Table 6-4.** Table 2. Catalytic activity and selectivity of Mb(H64V,V68A) and its stapled variants for cyclopropanation of styrene (1a) and pentafluorostyrene (1b) with ethyl 2-diazo-acetate (2) in the absence and in the presence of organic cosolvents



Entry	Variant	Product	Solvent	Yield, %	TON	% <i>de</i>	% <i>ee</i>
1	Mb(H64V,V68A)	<b>3a</b>	—	94	4,710 ± 180	99.4	99.1
2	sMb2	<b>3a</b>	—	94	4,740 ± 100	99.3	99.2
3	sMb5	<b>3a</b>	—	94	4,690 ± 25	99.1	99.1
4	sMb10	<b>3a</b>	—	94	4,710 ± 250	99.4	99.3
5	sMb13	<b>3a</b>	—	92	4,610 ± 140	99.2	99.2
6	Mb(H64V,V68A)	<b>3a</b>	30% EtOH	48	2,420 ± 280	98	92
7	sMb10	<b>3a</b>	30% EtOH	82	4,120 ± 270	99.3	97
8 <sup>±</sup>	sMb10	<b>3a</b>	30% DMF	84 (64)	4,220 ± 140 (3,480)	99.8 (99)	99.6 (99)
9 <sup>±</sup>	sMb13	<b>3a</b>	30% THF	81 (92)	4,360 ± 90 (4,660)	97 (99)	95 (94)
10 <sup>±</sup>	sMb10	<b>3a</b>	30% DMSO	64 (60)	3,220 ± 280 (3,000)	99.8 (99.8)	99.6 (99.4)
11 <sup>±</sup>	sMb10	<b>3a</b>	45% DMSO	39 (20)	1,990 ± 190 (1,020)	99.9 (99.9)	98.5 (98.5)
12	Mb(H64V,V68A)	<b>3b</b>	45% DMSO	54	270 ± 5	99.8	99.9
13	sMb10	<b>3b</b>	45% DMSO	80	395 ± 5	99.8	99.9
14	sMb13	<b>3b</b>	45% DMSO	77	380 ± 5	99.8	99.9

#### 4.4.7 Increased Stability Against Chemical Denaturation.

Enhanced robustness to chemical denaturation and organic solvents is a highly desirable trait of enzymes to be employed for synthetic applications.<sup>44, 45</sup> To examine this aspect, the most promising sMb constructs (sMb5, sMb10, and sMb13) were subjected to denaturation experiments in the presence of guanidinium chloride (Gnd·HCl) (SI Appendix, Fig. S9). Reflecting the trend in the thermostability assays, a progressive stability increase in the presence of the chaotropic agent was observed going from the singly stapled sMb5 ( $C_m$ : 1.79 M) to the doubly stapled sMb10 and sMb13 ( $C_m$ : 2.01 M and 2.08 M, respectively), compared with the parent enzyme ( $C_m$ : 1.55 M).

Next, we investigated the effect of the staples toward improving the performance of the sMb variants in organic solvents. In the presence of 30% (vol/vol) ethanol, Mb(H64V,V68A) catalyzes the cyclopropanation of styrene with reduced activity (48% yield; 2,420 TON; 51% relative activity) and lower enantioselectivity (92% *ee*) compared to the same reaction in buffer (Fig. 2E-F; entry 6 vs. 1, Table 2). Compared with Mb(H64V,V68A), both sMb10 and sMb13 better tolerate the presence of the organic cosolvent, affording the cyclopropanation product **3a** in higher yields (78–82%; 3,900–4,100 TON; entry 7, Table 2) and with higher diastereo- (1:400 vs. 1:130 diastereomeric ratio for *trans:cis*) and enantioselectivity (97% vs. 92% *ee*) (entry 7 vs. 6, Fig. 2F and Table 2). As stereoselectivity is highly sensitive to structural perturbation within the active site of these biocatalysts,<sup>30</sup> these results suggest that stabilization of the Mb scaffold by covalent stapling minimizes the disruptive effects induced by the organic solvent. A similar trend was observed upon comparing the activity and selectivity of

sMb5, sMb10, and sMb13 with those of Mb(H64V,V68A) in cyclopropanation reactions in the presence of high concentrations (30% vol/vol) of other organic solvents such as methanol, dimethylformamide (DMF), and DMSO (SI Appendix, Figs. S10-S12). In comparison, tetrahydrofuran (THF) was tolerated equally well by both the parent protein and the stapled variants (entry 9, Table 2). In the presence of these organic cosolvents, both sMb10 and sMb13 can produce **3a** in high yields (64–88%; 3,200–4,440 TON) and high stereoselectivity (99.5–99.8% *de*; 99.2–99.6% *ee*) (entries 8–10, Table 2). Furthermore, sMb10 and sMb13 maintain good catalytic activity (1,080–1,990 TON) as well as excellent stereoselectivity (>99% *de*; 97–98.5% *ee*) even in the presence of a nearly 1:1 mixture (45% vol/vol) of the buffer with DMF or DMSO (entry 11, Table 2 and SI Appendix, Figs. S11-S12). These results are remarkable considering that most heme-dependent enzymes are readily inactivated by low concentrations (>5–10% vol/vol) of these organic solvents.<sup>46,47</sup> Given their enhanced performance at high DMSO concentrations, sMb10 and sMb13 could be applied to afford the cyclopropanation of a water-insoluble substrate, that is, pentafluorostyrene (**1b**), with higher efficiency than possible using the parent Mb variant (Table 2, entries 13 and 14 vs. 12). Altogether, these results demonstrate the value of R-GPS toward enhancing the robustness and performance of a biocatalyst in the presence of chemical denaturants and organic cosolvents.

## 4.5 Discussion

As demonstrated through the design and characterization of highly stabilized variants of an Mb-based cyclopropanase, R-GPS enables the rapid identification of optimal sites

within a target protein scaffold for structural stabilization via "stapling" through nonnative covalent bridges. In its current implementation, R-GPS utilizes a bioorthogonal reaction between cysteine and the genetically encodable noncanonical amino acid O2beY to produce chemically stable and nonreducible thioether cross-links. Key advantages of this chemistry include its chemoselectivity, spatially controlled reactivity, and efficient and spontaneous formation of the cross-link at the posttranslational level and in the intracellular space. The latter feature allows for the production of the stapled proteins inside cells and eliminates the need for protein manipulation during or after purification.<sup>48</sup> Whereas R-GPS has been validated using the O2beY/Cys cross-linking method, we expect this approach can be extended to other noncanonical amino acids and/or other chemistries useful for protein cross-linking.<sup>49-52</sup>

Two computationally designed staples could be effectively combined in the final designs (sMb10 and sMb13) to achieve additive effects for stabilization against thermal and chemical denaturation. While the introduction of computationally designed nonnative disulfide bridges may result in misfolded variants due to disulfide scrambling,<sup>27</sup> no evidence of incorrect cross-linking was noted for the sMb variants containing the two thioether staples. We attribute this result to the heterotypic nature of the O2beY/Cys thioether linkage, which reduces the number of wrong cross-links that can be formed compared with the homotypic disulfide bridge. In addition, the presence of specific geometric requirements for the successful formation of the O2beY/Cys cross-link, as revealed by the present studies (SI Appendix, Fig. S1), likely contributes to further disfavor the formation of incorrect linkages. At the same time, a noticeable reduction in the protein yields was observed upon the incorporation of multiple copies of O2beY via

amber stop codon suppression [2 mg/L culture for sMb10 compared with 10–15 mg/L for sMb2/sMb5 and 30 mg/L for Mb(H64V,V68A)]. While this result is not surprising given the less-efficient incorporation of noncanonical amino acids (ncAAs) compared with natural amino acids, various strategies have recently enabled the expression of proteins containing multiple ncAAs,<sup>53-55</sup> and these protocols should increase the accessibility of proteins containing multiple staples designed with the present methodology.

R-GPS was able to identify a small set of structural modifications (i.e., four and five mutations for sMb10 and sMb13, respectively) capable of conferring these carbene transferase enzymes significantly increased stability against thermal and chemical denaturation without impairing their catalytic activity and subtle functional properties such as stereoselectivity. The improved stability properties of these biocatalysts translated into improved performance and stereoselectivity in reactions conducted in high concentrations of organic solvents, also enabling the more efficient transformation of a water-insoluble olefin substrate. Since R-GPS involves a small set of amino acid substitutions, most of the protein residues remain available for stabilization based on the optimization of interactions in the native state. Thus, it should be possible to combine R-GPS with other computational design<sup>13, 14, 17</sup> and/or experimental methods<sup>4-6</sup> to achieve additive or synergistic effects in protein/enzyme stabilization. The minimally invasive approach to protein stabilization described here is expected to provide a valuable strategy for greatly enhancing the stability of enzymes and proteins without impacting their functional properties.

## 4.6 References

1. Bornscheuer UT, et al.: **Engineering the third wave of biocatalysis.** *Nature* 2012, **485**:185–194.
2. Clouthier CM, Pelletier JN: **Expanding the organic toolbox: A guide to integrating biocatalysis in synthesis.** *Chem Soc Rev* 2012, **41**:1585–1605.
3. Bloom JD, Labthavikul ST, Otey CR, Arnold FH: **Protein stability promotes evolvability.** *Proc Natl Acad Sci USA* 2006, **103**:5869–5874.
4. Gershenson A, Arnold FH. **Enzyme stabilization by directed evolution.** *Genet Eng (N Y)* 2000, **22**:55–76.
5. Lehmann M, Wyss M: **Engineering proteins for thermostability: The use of sequence alignments versus rational design and directed evolution.** *Curr Opin Biotechnol* 2001, **12**:371–375.
6. Bommarius AS, Paye MF: **Stabilizing biocatalysts.** *Chem Soc Rev* 2013, **42**:6534–6565.
7. Tracewell CA, Arnold FH: **Directed enzyme evolution: Climbing fitness peaks one amino acid at a time.** *Curr Opin Chem Biol* 2009, **13**:3–9.
8. Magliery TJ: **Protein stability: Computation, sequence statistics, and new experimental methods.** *Curr Opin Struct Biol* 2015, **33**:161–168.

9. Malakauskas SM, Mayo SL: **Design, structure and stability of a hyperthermophilic protein variant.** *Nat Struct Biol* 1998, **5**:470–475.
10. Korkegian A, Black ME, Baker D, Stoddard BL: **Computational thermostabilization of an enzyme.** *Science* 2005, **308**:857–860.
11. Borgo B, Havranek JJ: **Automated selection of stabilizing mutations in designed and natural proteins.** *Proc Natl Acad Sci USA* 2012, **109**:1494–1499.
12. Li Y, et al.: **A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments.** *Nat Biotechnol* 2007, **25**:1051–1056.
13. Goldenzweig A, et al.: **Automated structure- and sequence-based design of proteins for high bacterial expression and stability.** *Mol Cell* 2016, **63**:337–346.
14. Bednar D, et al.: **FireProt: Energy- and evolution-based computational design of thermostable multiple-point mutants.** *PLOS Comput Biol* 2015, **11**:e1004556.
15. Gribenko AV, et al.: **Rational stabilization of enzymes by computational redesign of surface charge-charge interactions.** *Proc Natl Acad Sci USA* 2009, **106**:2601–2606.
16. Bjørk A, Dalhus B, Mantzilas D, Sirevåg R, Eijsink VG: **Large improvement in the thermal stability of a tetrameric malate dehydrogenase by single point mutations at the dimer-dimer interface.** *J Mol Biol* 2004, **341**:1215–1226.

17. Wijma HJ, et al.: **Computationally designed libraries for rapid enzyme stabilization.** *Protein Eng Des Sel* 2014, **27**:49–58.
18. Van den Burg B, Vriend G, Veltman OR, Venema G, Eijsink VG: **Engineering an enzyme to resist boiling.** *Proc Natl Acad Sci USA* 1998, **95**:2056–2060.
19. Zou J, Song B, Simmerling C, Raleigh D: **Experimental and computational analysis of protein stabilization by Gly-to-d-Ala substitution: A convolution of native state and unfolded state effects.** *J Am Chem Soc* 2016, **138**:15682–15689.
20. Dombkowski AA, Sultana KZ, Craig DB: **Protein disulfide engineering.** *FEBS Lett.* 2014, **588**:206–212.
21. Matsumura M, Signor G, Matthews BW: **Substantial increase of protein stability by multiple disulphide bonds.** *Nature* 1989, **342**:291–293.
22. Liu T, et al.: **Enhancing protein stability with extended disulfide bonds.** *Proc Natl Acad Sci USA* 2016, **113**:5910–5915.
23. Sanchez-Romero I, et al.: **Mechanism of protein kinetic stabilization by engineered disulfide crosslinks.** *PLoS One* 2013, **8**:e70013.
24. Tyagi V, Bonn RB, Fasan R: **Intermolecular carbene S-H insertion catalysed by engineered myoglobin-based catalysts.** *Chem Sci (Camb)* 2015, **6**:2488–2494.

25. Carninci P, et al.: **Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA.** *Proc Natl Acad Sci USA* 1998, **95**:520–524.
26. Tyagi V, Fasan R: **Myoglobin-catalyzed olefination of aldehydes.** *Angew Chem Int Ed Engl* 2016, **55**:2512–2516.
27. Tanghe M, et al.: **Disulfide bridges as essential elements for the thermostability of lytic polysaccharide monooxygenase LPMO10C from *Streptomyces coelicolor*.** *Protein Eng Des Sel* 2017, **30**:401–408.
28. Bionda N, Cryan AL, Fasan R: **Bioinspired strategy for the ribosomal synthesis of thioether-bridged macrocyclic peptides in bacteria.** *ACS Chem Biol* 2014, **9**:2008–2013.
29. Bordeaux M, Tyagi V, Fasan R: **Highly diastereoselective and enantioselective olefin cyclopropanation using engineered myoglobin-based catalysts.** *Angew Chem Int Ed Engl* 2015, **54**:1744–1748.
30. Bajaj P, Sreenilayam G, Tyagi V, Fasan R: **Gram-scale synthesis of chiral cyclopropane-containing drugs and drug precursors with engineered myoglobin catalysts featuring complementary stereoselectivity.** *Angew Chem Int Ed Engl* 2016, **55**:16110–16114.

31. Tyagi V, Sreenilayam G, Bajaj P, Tinoco A, Fasan R: **Biocatalytic synthesis of allylic and allenyl sulfides through a myoglobin-catalyzed Doyle-Kirmse reaction.** *Angew Chem Int Ed Engl* 2016, **55**:13562–13566.
32. Bionda N, Fasan R: **Ribosomal synthesis of natural-product-like bicyclic peptides in Escherichia coli.** *ChemBioChem* 2015, **16**:2011–2016.
33. Betz SF: **Disulfide bonds and the stability of globular-proteins.** *Protein Sci* 1993, **2**:1551–1558.
34. Zhang T, Bertelsen E, Alber T: **Entropic effects of disulphide bonds on protein stability.** *Nat Struct Biol* 1994, **1**:434–438.
35. Flory PJ: **Theory of elastic mechanisms in fibrous proteins.** *J Am Chem Soc* 1956, **78**:5222–5234.
36. Chan HS, Dill KA: **Intrachain loops in polymers—Effects of excluded volume.** *J Chem Phys* 1989, **90**:492–509.
37. Zanghellini A, et al.: **New algorithms and an in silico benchmark for computational enzyme design.** *Protein Sci* 2006, **15**:2785–2794.
38. Kuhlman B, Baker D: **Native protein sequences are close to optimal for their structures.** *Proc Natl Acad Sci USA* 2000, **97**:10383–10388, and erratum (2000) 97:13460.

39. Liu CC, Schultz PG: **Adding new chemistries to the genetic code.** *Annu Rev Biochem* 2010, **79**:413–444.
40. Deiters A, Schultz PG: **In vivo incorporation of an alkyne into proteins in *Escherichia coli*.** *Bioorg Med Chem Lett* 2005, **15**:1521–1524.
41. Nagatani RA, Gonzalez A, Shoichet BK, Brinen LS, Babbitt PC: **Stability for function trade-offs in the enolase superfamily "catalytic module".** *Biochemistry* 2007, **46**:6688–6695.
42. Hamamatsu N, et al.: **Directed evolution by accumulating tailored mutations: Thermostabilization of lactate oxidase with less trade-off with catalytic activity.** *Protein Eng Des Sel* 2006, **19**:483–489.
43. Zhong CQ, et al.: **Improvement of low-temperature caseinolytic activity of a thermophilic subtilase by directed evolution and site-directed mutagenesis.** *Biotechnol Bioeng* 2009, **104**:862–870.
44. Savile CK, et al.: **Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture.** *Science* 2010, **329**:305–309.
45. Fox RJ, et al.: **Improving catalytic function by ProSAR-driven enzyme evolution.** *Nat Biotechnol* 2007, **25**:338–344.
46. Wong TS, Arnold FH, Schwaneberg U: **Laboratory evolution of cytochrome p450 BM-3 monooxygenase for organic cosolvents.** *Biotechnol Bioeng* 2004, **85**:351–358.

47. Chauret N, Gauthier A, Nicoll-Griffith DA: **Effect of common organic solvents on in vitro cytochrome P450-mediated metabolic activities in human liver microsomes.** *Drug Metab Dispos* 1998, **26**:1–4.
48. Abdeljabbar DM, Piscotta FJ, Zhang S, James Link A: **Protein stapling via azide-alkyne ligation.** *Chem Commun (Camb)* 2014, **50**:14900–14903.
49. Furman JL, et al.: **A genetically encoded aza-Michael acceptor for covalent cross-linking of protein-receptor complexes.** *J Am Chem Soc* 2014, **136**:8411–8417.
50. Xiang Z, et al.: **Proximity-enabled protein crosslinking through genetically encoding haloalkane unnatural amino acids.** *Angew Chem Int Ed Engl* 2014, **53**:2190–2193.
51. Chen XH, et al.: **Genetically encoding an electrophilic amino acid for protein stapling and covalent binding to native receptors.** *ACS Chem Biol* 2014, **9**:1956–1961.
52. Xuan W, Shao S, Schultz PG: **Protein crosslinking by genetically encoded noncanonical amino acids with reactive aryl carbamate side chains.** *Angew Chem Int Ed Engl* 2017, **56**:5096–5100.
53. Guo J, Melançon CE, Lee HS, Groff D, Schultz PG: **Evolution of amber suppressor tRNAs for efficient bacterial production of proteins containing nonnatural amino acids.** *Angew Chem Int Ed Engl* 2009, **48**:9148–9151.

54. Johnson DBF, et al.: **RF1 knockout allows ribosomal incorporation of unnatural amino acids at multiple sites.** *Nat Chem Biol* 2011, **7**:779–786.
55. Lajoie MJ, et al.: **Genomically recoded organisms expand biological functions.** *Science* 2013, **342**:357–360.

## 4.7 Supplementary Information

### 4.7.1 Supplemental Text

#### 4.7.1.1 Near-attack conformation (NAC) analysis.

The positional dependence of the stapling reaction could be rationalized based on analysis of the rotamers of Cys and O2beY residues (Figure 1B) using models of the sMb constructs in their corresponding unstapled form. Since the initial designs were based solely on the compatibility of the thioether product with the chosen residue positions in the native structure, the accessibility of near-attack conformations (NACs) compatible with a bimolecular nucleophilic substitution reaction (SN2) between the thiol group of the cysteine and the alkyl bromide group in O2beY (Figure S1A) was assessed for each of these constructs. These analyses showed that SN2- compatible NACs are accessible to sMb2, sMb3, sMb5, and sMb7 (Table 1; Figures S1B-D and S2), all of which were found to be crosslinked experimentally. In contrast, the most favourable conformations for the unstapled forms of sMb1, sMb6, sMb8 and sMb9 involve a syn positioning of the thiolate with respect to the bromide leaving group ( $\theta_{\text{attack}} < 110\text{-}130^\circ$ ; Table 1), which is incompatible with the preferred trajectory of nucleophile attack in SN2 reactions (Figure S1A). These results are consistent with the lack of crosslinking observed for these

constructs. The unstapled sMb4 model was also found to favour an SN2-incompatible conformation for the Cys and O2beY residues ( $\theta_{\text{attack}} = 98^\circ$ ), but experimental characterization showed that this construct is able to undergo partial crosslinking, possibly due to stabilization provided to the developing negative charge on the leaving group bromide by a nearby Arg31 residue (Figure S1D), which is reminiscent of halide stabilization mechanism in some dehalogenase enzymes (1). Finally, while SN2-compatible NACs were accessible for many of the constructs containing a solvent-exposed crosslink, neither of the two constructs containing a buried crosslinked (sMb6 and sMb8) were found capable of accessing these productive conformations, primarily due to steric clashes at the bromine atom within the more confined protein microenvironment (Figure S2A-B). Overall, the good agreement between NAC analysis and experimental results supports the feasibility of this 3 modeling approach for designing Cys/O2beY substitutions for productive formation of thioether staples.

#### 4.7.1.2 Effect of F106A mutation.

A F106A substitution was introduced in the sMb10-derived construct sMb11 with the goal of testing the contribution of a potential  $\pi$ - $\pi$  stacking interaction between the O2beY residue in position 36 and the nearby Phe106 residue observed in design models of sMb2 and sMb10 (Figure S3). This interaction mimics the  $\pi$ - $\pi$  stacking arrangement observed for the side chains of the highly conserved His36 and Phe106 residues in wild-type myoglobins from several species and observed in both crystallographic and NMR structures (2). Interestingly, the F106A substitution led to a marked decrease in stability (T50) compared to sMb10 (Table 1), highlighting the importance of the computed

interaction. We confirmed experimentally that both crosslinks are formed in sMb11 (Figure S6), indicating that the decrease in stability exhibited by sMb11 compared to sMb10 can be ascribed directly to the F106A substitution. These results indicate that, in line with our second design criterion, energetics of inter-residue interactions in the microenvironment of the designed crosslinks can significantly influence their stabilization effect.

#### 4.7.2 Materials and Methods

##### 4.7.2.1 Computational Design

RosettaMatch (3) was used to determine suitable locations for placement of the O2beY/Cys staples. To model the stapled state (Figure 1A), ethyl group was considered as a separate ligand and deprotonated tyrosine and cysteine residues were used to place the ethyl group in a compatible geometry for the formation of O—C and S—C bonds, respectively (Figure 1B). These compatible geometries were derived from the Cambridge Structural Database. Geometrically compatible staple conformations were accepted if the two interacting residues were positioned in the first and last 30% of the sequence of myoglobin scaffold (PDB code: 1JP9), respectively. The accepted stapled models were subjected to a Rosetta EnzymeDesign (4) protocol for identifying stabilizing substitutions. As we aimed for designs with a minimal number of substitutions, visual inspection was used to revert Rosetta-suggested substitutions distal from the crosslinking site, when present. The resulting structures were subject to an additional round of energy minimization to calculate energies. All Rosetta files required to perform simulations are provided as Supplementary Information.

#### 4.7.2.2 Near-attack conformation (NAC) analysis

For NAC analysis in the sMb1-9 models, O2beY and Cys were modeled in the pre-attack conformations. The bond angle and lengths for O2beY were derived from the Cambridge Structural Database. Geometric constraints for the bimolecular nucleophilic substitution (SN2) reaction between Cys and O2beY residues were defined by specifying optimal values for the angle of nucleophilic attack ( $\theta_{\text{attack}}$ ) and dihedral angle  $\chi_4$  as  $180^\circ$ . This set of geometric constraints was applied in conjunction with Rosetta energy minimization and rotameric sampling to obtain pre-attack sMb models. After minimization, each sMb model was ranked with respect to the others by deviation from geometric constraints and Rosetta residue energy scores (Table S2) All files required to perform NAC analysis are provided as separate SI Appendix material.

#### 4.7.2.3 Cloning

The genes encoding for the Mb variants were cloned by overlap extension PCR (OE-PCR) using a previously described pET22-based vector containing the Mb(H64V,V68A) gene under a IPTG inducible T7 promoter(5) as the template and the oligonucleotides listed in Table S1 as the primers. After amplification, the OE-PCR product (512 bp) was inserted into the Xba I/Xho I cassette of a pET22 vector. After ligation, the recombinant clones were selected on LB agar plates containing ampicillin (100 mg/L) and confirmed by DNA sequencing.

#### 4.7.2.4 Protein Expression

For protein expression, the pET22-based plasmid encoding for the Mb variant was co-transformed into BL21(DE3) cells along with a pEVOL-based vector encoding for the O2beY-specific orthogonal AARS/tRNA pair.<sup>(6)</sup> Recombinant cells were selected on LB agar plates containing ampicillin (100 mg/L) and chloramphenicol (34 mg/L) and grown in LB medium containing the same antibiotics overnight at 37°C. Overnight cultures were used to inoculate 1 L M9 minimal medium containing 0.5% (w/v) yeast extract, 1% (v/v) glycerol, ampicillin (100 mg/L), and chloramphenicol (34 mg/L). Cell cultures were grown at 37°C until the OD<sub>600</sub> reached 0.6-0.8, followed by 5- fold condensation in the same medium. Expression of OpgY2-RS was induced with 4 mM L-arabinose (600 mg/L), followed by addition of O2beY at 1 mM final concentration. Expression of the myoglobin variant was induced with 0.5 mM IPTG followed by incubation at 27°C with shaking (180 rpm) for 16-20 hours. The OpgY-containing variants sMb2(OpgY) and sMb5(OpgY) were expressed in a similar manner with the difference that an OpgY-specific orthogonal AARS/tRNA pair was used and OpgY was added to the culture medium.<sup>(6)</sup>

#### 4.7.2.5 Protein Purification and Characterization

After harvesting by centrifugation, cells were resuspended in Ni-NTA lysis buffer (50 mM KPi, 250 mM NaCl, 10 mM histidine, pH 8.0) and lysed by sonication. The clarified lysate was loaded on a Ni-NTA column equilibrated with Ni-NTA Lysis Buffer. The resin was washed with 50 mL Ni-NTA Lysis Buffer and with 50 mL of Ni-NTA Wash Buffer (50 mM KPi, 250 mM NaCl, 20 mM histidine, pH 8.0). Proteins were eluted with

Ni-NTA Elution Buffer (50 mM KPi, 250 mM NaCl, 250 mM Histidine, pH 7.0), and then concentrated and buffer exchanged with 50 mM potassium phosphate buffer (pH 7.0) using 10 kDa Centricon filters. The proteins were purified using gel filtration chromatography using a Superdex 75 10/300 GL column and isocratic elution in 50 mM potassium phosphate buffer (pH 7.0) at 1.0 mL/min. The concentrations of the Mb variants were determined based on the absorption at the Soret band ( $\epsilon_{409} = 156,000 \text{ M}^{-1} \text{ cm}^{-1}$ ). Protein masses were analyzed using a Bruker AutoFlex II MALDI-TOF MS spectrometer using sinapinic acid as matrix. Near-UV circular dichroism spectra (250-190 nm) were obtained using 3  $\mu\text{M}$  solutions of purified Mb variant in 50 mM potassium phosphate buffer (pH 7.0) and recorded at 20°C at a scan rate of 50 nm/min with a bandwidth of 1 nm and an averaging time of 10 seconds per measurement. The raw signal ( $\theta_d$ , mDeg) was background subtracted against buffer and converted to molar residue ellipticity ( $\theta_{\text{MRE}}$ ) using  $\theta_{\text{MRE}} = \theta_d / (c \ln R)$ , where  $c$  is the concentration (M),  $l$  is the path length (1 mm), and  $nR$  is the number of residues in the protein.

#### 4.7.2.6 $T_m$ determination

Thermal denaturation experiments were carried out using a JASCO J1100 CD spectrophotometer equipped with variable temperature/wavelength denaturation analysis software and samples of purified Mb variant at 3  $\mu\text{M}$  in 50 mM potassium phosphate buffer (pH 7.0). Thermal denaturation curves were measured by monitoring the change in molar ellipticity at 222 nm ( $\theta_{222}$ ) over a temperature range from 20°C to 100°C. The temperature increase was set to 0.5°C per minute with an equilibration time of 10 seconds. Data integration time for the melt curve was set to 4 seconds with a bandwidth

of 1 nm. Linear baselines for the folded ( $\theta_f$ ) and unfolded state ( $\theta_u$ ) were generated using the low temperature ( $\theta_f = m_f T + b_f$ ) and high temperature ( $\theta_u = m_u T + b_u$ ) equations fitted to the experimental data before and after global unfolding, respectively (Figure S7). Using these equations, the melt data were converted to fraction of folded protein ( $F_f$ ) vs. temperature plots and the resulting curve was fitted to a sigmoidal equation ( $\theta_{fit}$ ) via non-linear regression analysis in SigmaPlot (Figure S7), from which apparent melting temperatures were derived. The reported mean values and standard errors were derived from experiments performed at least in duplicate.

#### 4.7.2.7 T50 analysis

For the thermal stability experiments (T50 determination), 500  $\mu$ L of protein solution at 3.5  $\mu$ M in 50 mM potassium phosphate buffer (pH 7.0) were incubated for 10 minutes at varying temperature between 20°C and 95°C (10°C intervals) in a thermoblock. After incubation, the protein samples were centrifuged (14,000 rpm, 4°C, 10 minutes) and the supernatant was transferred to a 96 well plate. Visible spectra were recorded between 300 and 500 nm using a Tecan X microtiter plate reader. The residual fraction of holoprotein in each sample was determined based on the intensity of the Soret band (410 nm) after normalization to the sample incubated at 20°C. Half-maximal denaturation temperatures (T50) were calculated from the fraction of folded protein vs. temperature plots by fitting the data to a three-parameter sigmoidal equation in SigmaPlot. The reported mean values and standard errors were derived from experiments performed at least in duplicate.

#### 4.7.2.8 Chemical Denaturation Experiments

Chemical denaturation curves were measured via circular dichroism by monitoring the change in molar ellipticity at 220 nm ( $\theta_{220}$ ) for the Mb variants in the presence of varying concentration of guanidinium hydrochloride (Gdn·HCl). Protein solutions containing 3  $\mu$ M Mb variant in 50 mM phosphate buffer (pH 7.0) and between 0 and 5.5 M Gdn·HCl in 0.25-0.5 M increments were allowed to equilibrate for 30 minutes at 20°C with shaking (700 rpm) followed by centrifugation at 14,000 rpm at room temperature for 10 minutes. CD spectra of the protein in the supernatant were recorded from 190 to 250 nm and absorbance data ( $\theta_{220}$ ) were normalized to the signal corresponding to the reference sample with no Gdn·HCl (set as 1) and the sample with 5.5 M Gdn·HCl (set as 0). The resulting fraction of folded protein (f) vs. [Gdn·HCl] curves were fitted to a sigmoidal curve equation via non-linear regression analysis in SigmaPlot, from which half-maximal denaturation concentration ( $C_m$ ) values were obtained. All measurements were performed at least in duplicate.

#### 4.7.2.9 Cyclopropanation Activity

The cyclopropanation reactions were performed at a 500  $\mu$ L scale using 2  $\mu$ M myoglobin, 10 mM styrene (1a), 20 mM ethyl  $\alpha$ -diazoacetate (EDA, 2), and 10 mM sodium dithionite. For the reaction setup, a solution containing sodium dithionite (10 mM in 50 mM potassium phosphate buffer, pH 7.0) was degassed by bubbling argon for 3 minutes in a sealed vial. For the stability tests in the presence of the organic cosolvents, the dithionite solution was added with 30 (or 45%) v/v of the appropriate organic solvent. In a separate sealed vial, a stock protein solution (2  $\mu$ M in 50 mM phosphate buffer, pH

7.0) was degassed in a similar manner. The two solutions were then mixed together via cannulation and reactions were initiated by addition of 10  $\mu$ L of 0.4 M styrene solution in ethanol, followed by the addition of 10  $\mu$ L of 0.8 M EDA solution. The reaction mixture was stirred overnight at room temperature and under positive argon pressure. The reactions with pentafluorostyrene (1b) were carried out in a similar manner but using 5 mM olefin, 10 mM EDA, and 10  $\mu$ M protein.

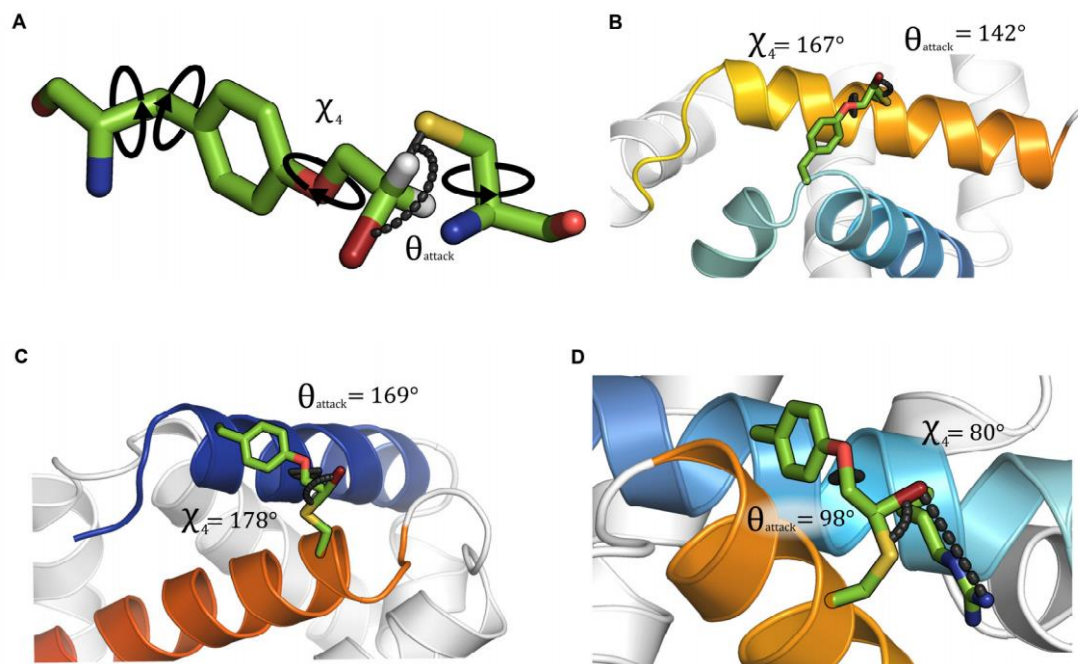
#### 4.7.2.10 Product analysis

The reactions were analyzed by adding 20  $\mu$ L of internal standard (benzodioxole, 100 mM in ethanol) to the reaction mixture, followed by extraction with 400  $\mu$ L of dichloromethane and analyzed by chiral gas chromatography as described.<sup>(5)</sup> Yields and number of turnovers (TON) were calculated based on the amounts of cyclopropane product as determined using calibration curve with authentic standards. All measurements were performed at least in duplicate. Diastereomeric and enantiomeric excess for compound 3a was determined based on the relative distribution of the four stereoisomer products as described previously.<sup>(5)</sup> Enantiomer resolution for compound 3b was performed by Supercritical Fluid Chromatography (SFC), using a JASCO Analytical and Semi-Preparative SFC instrument equipped with a Daicel Chiralpak IF column (0.46 cm  $\times$  25 cm), column oven (35  $^{\circ}$ C), photodiode array detector, a backpressure regulator (12.0 MPa), a carbon dioxide pump and a sample injection volume of 3  $\mu$ L. Samples were eluted using an isocratic solvent system with 100% liquid CO<sub>2</sub> at an elution rate of 4 mL/min and detected at  $\lambda$  = 220 nm. Total run time was 10.2 min.

## 4.7.2.11 Chemical synthesis of racemic standard for 3b

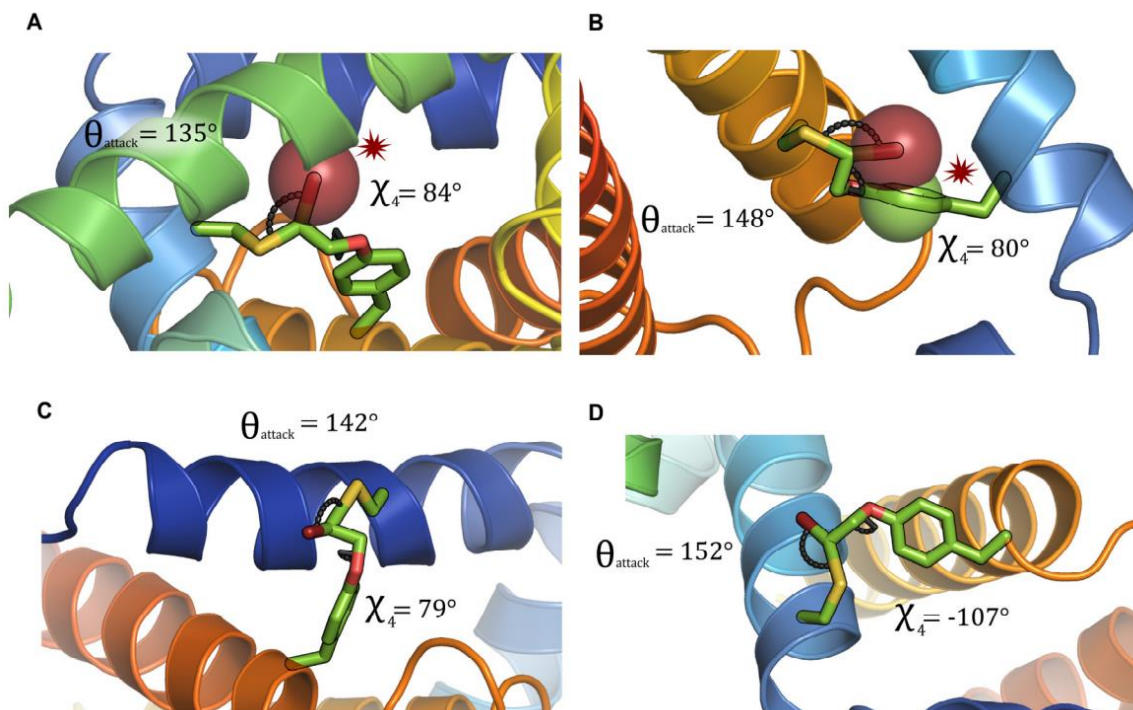
To a flame dried round bottom flask under argon, equipped with a stir bar was added pentafluorostyrene (250 mg, 1.29 mmol, 1 equiv.) and Rh<sub>2</sub>(OAc)<sub>4</sub> (5 mol%) in dry CH<sub>2</sub>Cl<sub>2</sub> (5 mL). To this solution was added dropwise a solution of EDA (1 equiv.) in dry CH<sub>2</sub>Cl<sub>2</sub> (1.5 mL) over 4 hours using a syringe pump at room temperature. The resulting mixture was stirred at 25°C overnight. The solvent was removed under reduced pressure and the crude mixture was purified by flash column chromatography (hexanes/CH<sub>2</sub>Cl<sub>2</sub> 75:25) to provide the cyclopropanation product 3b in 50% yield as a mixture of trans and cis isomers (2 : 1) both in racemic form. The trans isomer was isolated as a colorless liquid (50%) using flash chromatography with a step gradient from 10% to 30% CH<sub>2</sub>Cl<sub>2</sub> in hexanes. GC-MS m/z (% relative intensity): 280(52.8), 252(47.6), 235(68.4), 225(32.6), 207(79.8), 187(100.0), 181(76.5); <sup>1</sup>H NMR (CDCl<sub>3</sub>, 500 MHz): δ 4.19 (q, J = 7.0 Hz, 2H), 2.44 (m, 1H), 2.15 (m, 1H), 1.61 (m, 2H), 1.28 (t, J = 7.0 Hz, 3H); <sup>19</sup>F NMR (CDCl<sub>3</sub>, 400 MHz): δ -143.13 (t, 2F), -156.45 (m, 1F), -162.00 (m, 2F) ppm (reference: CF<sub>3</sub>Cl<sub>3</sub>).

## 4.7.3 Supplemental Figures



**Figure 15-4.** Figure S1. Near-attack conformation (NAC) analysis.

(A) Optimal geometry for backside attack of thiolate nucleophile (from Cys) on the halogen-bearing carbon of O2beY. The angle of nucleophilic attack ( $\theta_{\text{attack}}$ ; ideal value  $180^\circ$ ) and the terminal dihedral angle of O2beY ( $\chi_4$ ; ideal value  $180^\circ$ ) were used as parameters to assess accessibility of  $\text{S}_{\text{N}}2$ -compatible NACs. Calculated NACs for sMb2 (B), sMb5 (C) and sMb4 (D) are shown. In sMb4 NAC, a hydrogen bond between R31 and the leaving group bromide is observed.



**Figure 16-4.** Figure S2. NAC analysis for additional sMb variants.

In sMb6 and sMb8 (A-B), clashes between the bromide leaving group and neighbouring residues in the protein may disfavor staple formation, while in sMb3 and sMb7 (C-D), crosslinking is favored due to the accessibility of SN2-compatible geometries ( $\theta_{\text{attack}}$ ,  $\chi_4$ ) for nucleophilic attack of the cysteine side-chain thiol group on the alkyl bromide group of O2beY.

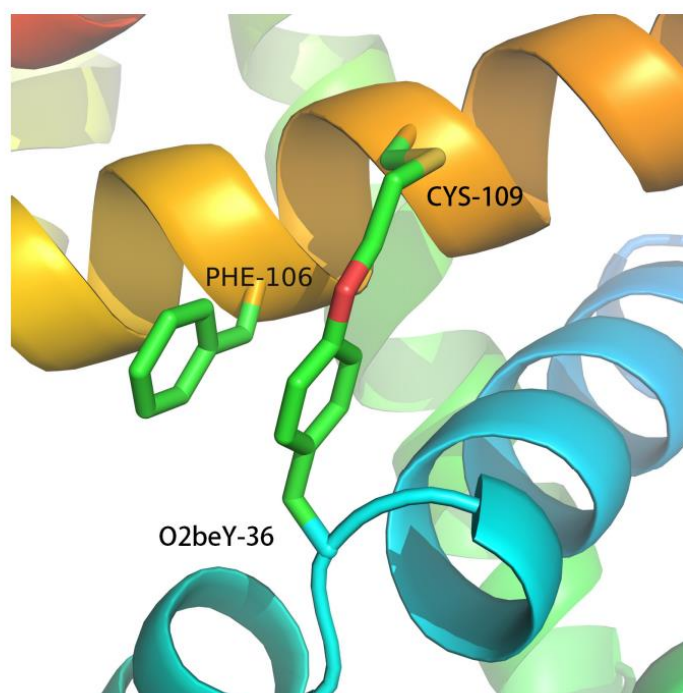
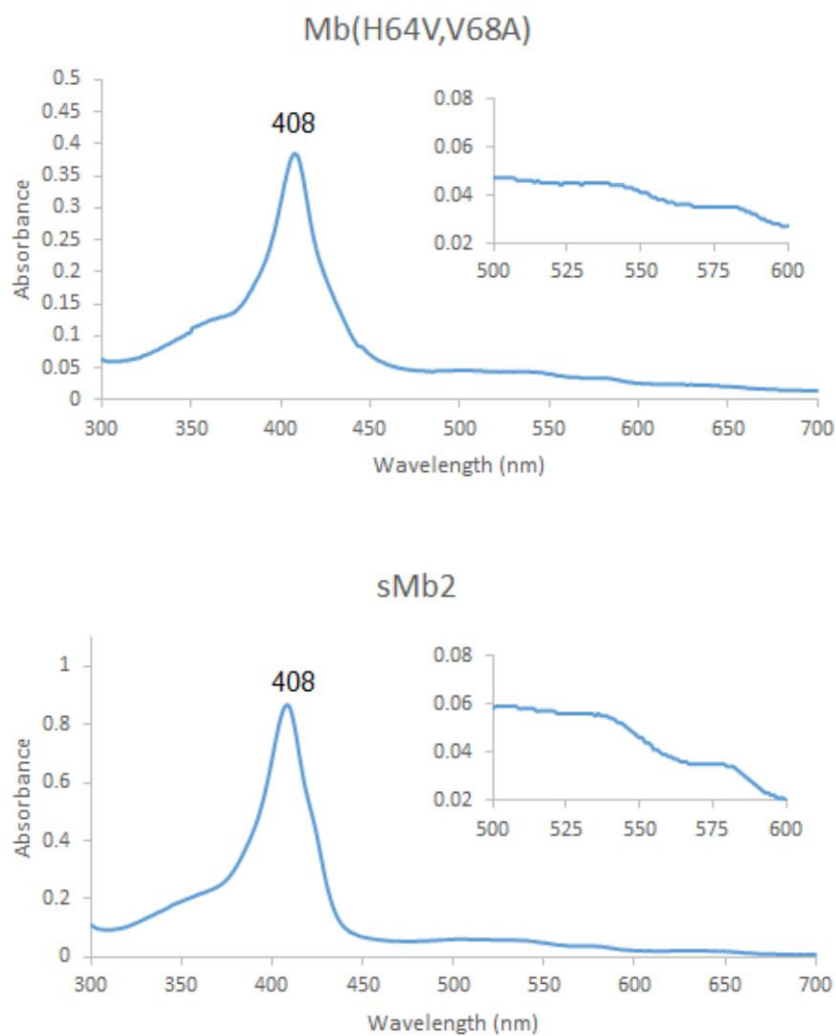


Figure 17-4. Figure S3. Interaction between O2beY36 and Phe106 in sMb2.

A parallel pi-pi stack interaction between the sidechains of Phe106 and O2beY36 was observed in in the design model of sMb2. Experimental characterization of the sMb11 variant, which carries a F106A substitution, suggest that this interaction contributes to the observed thermostabilization in constructs bearing the O2beY36/Cys109 staple (i.e., Mb2, sMb10, sMb13).



**Figure 18-4.** Figure S4. Visible-range electronic absorption spectra for Mb(H64V,V68A) and representative sMb variants in the ferric form.

The  $\lambda_{\text{max}}$  value for the characteristic Soret band is indicated. The insert show the Q band region of the proteins.

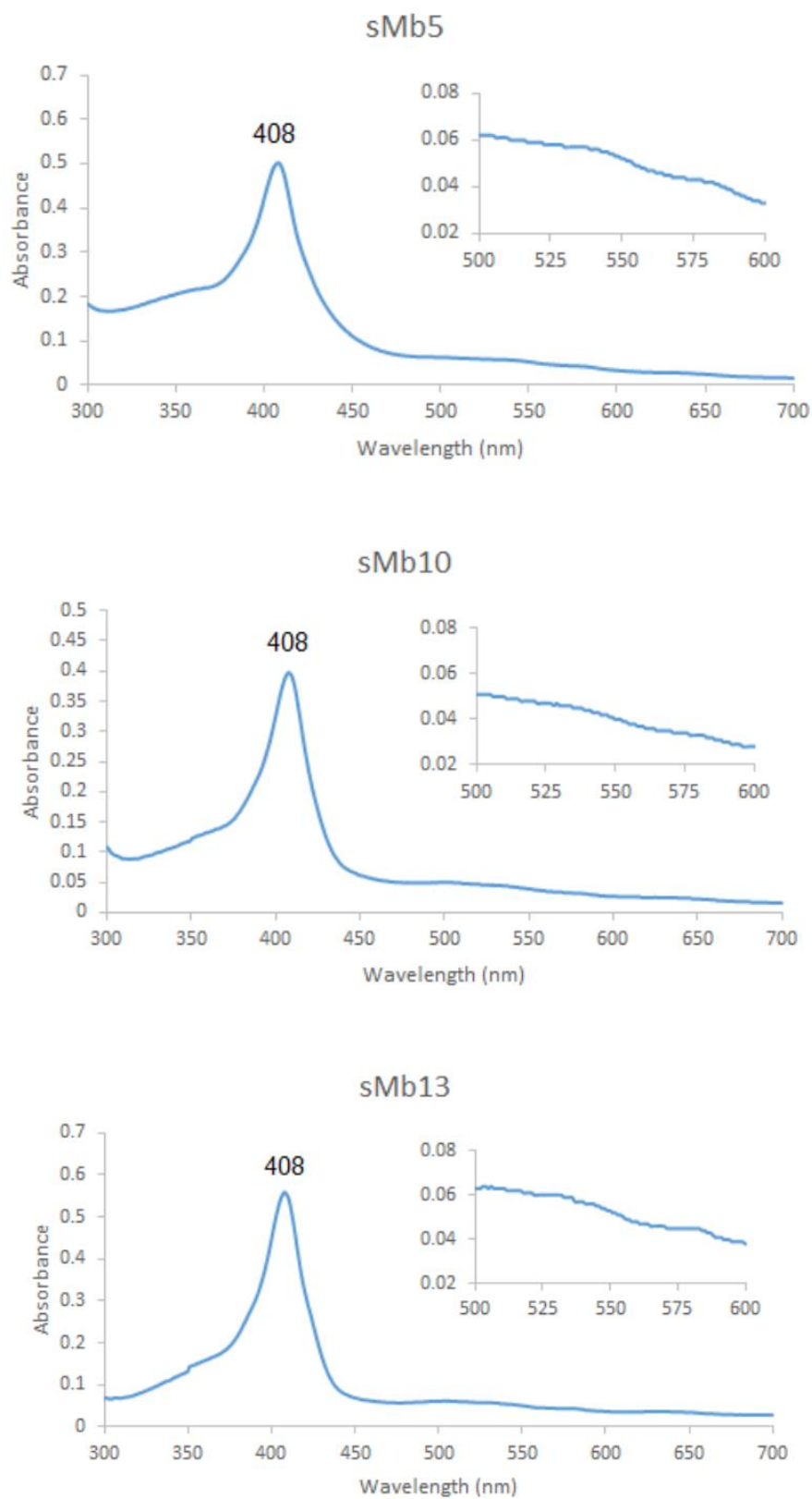
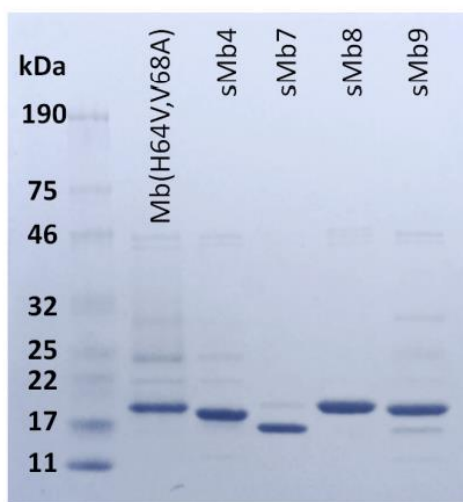


Figure 19-4. Figure S4. (cont.)



**Figure 20-4.** Figure S5. SDS-PAGE gel of remaining sMb variants not included in Figure 2A.

The gel shows increased electrophoretic mobility for sMb4 and sMb7 and similar mobility for sMb8 and sMb9 compared to the parent Mb(H64V,V68A), which is consistent with the presence and absence, respectively, of the O2beY/Cys crosslink in these proteins as determined by mass spectrometry.

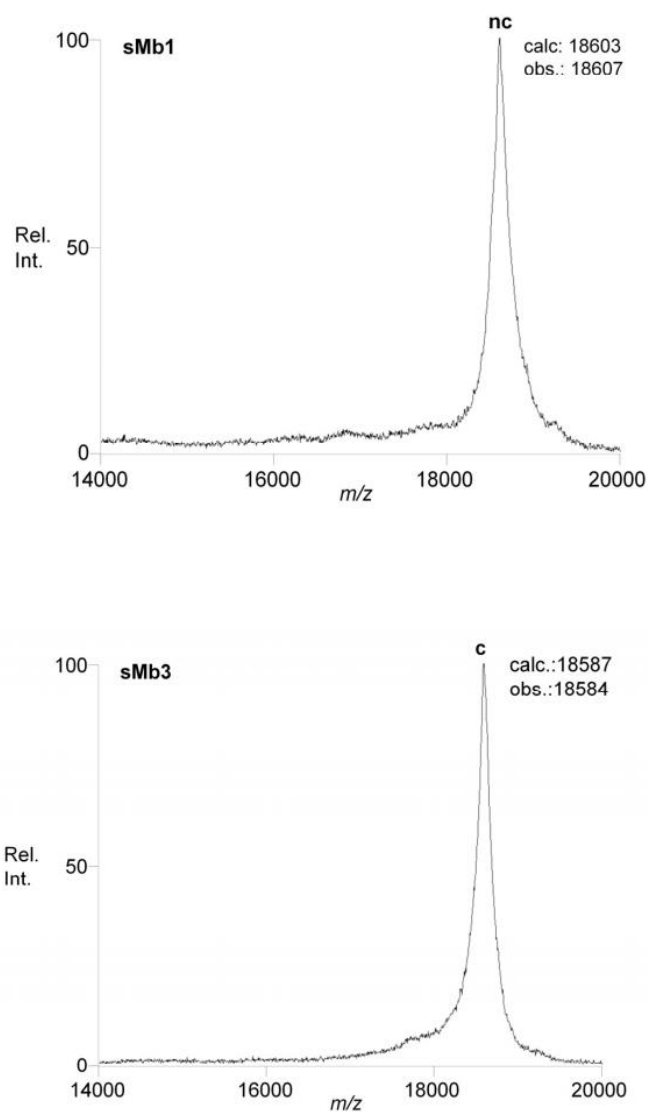


Figure 21-4. Figure S6. MALDI-TOF MS spectrum of additional sMb variants not included in Figure 2B.

Observed and calculated masses corresponding to the proton adduct ( $[M+H]^+$ ) of the protein species are indicated. 'c' = crosslinked; 'c(2x)' = doubly crosslinked; 'nc' = not crosslinked.

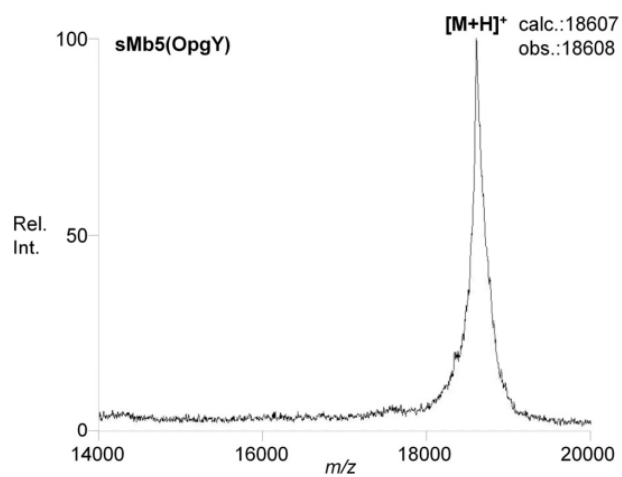
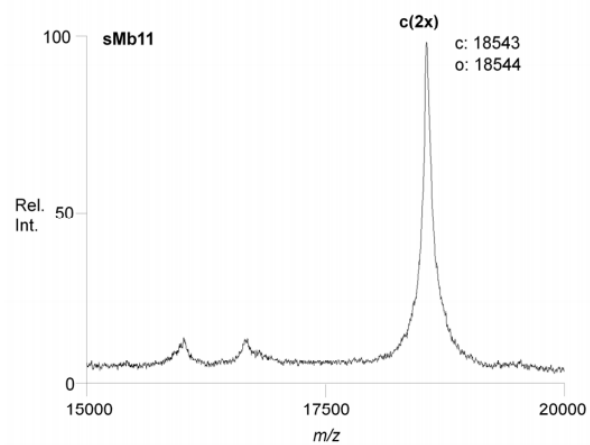
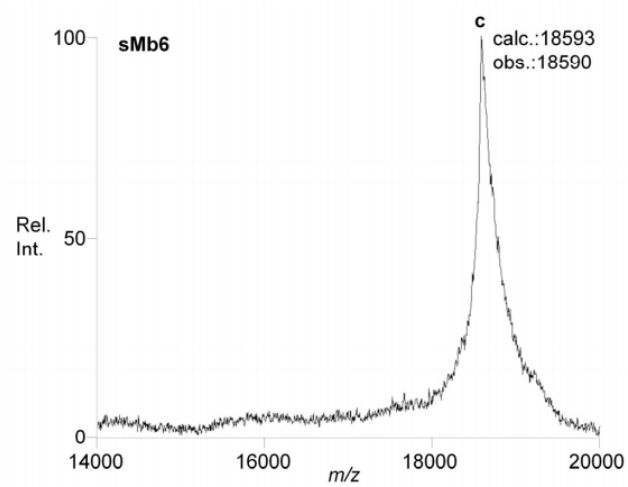
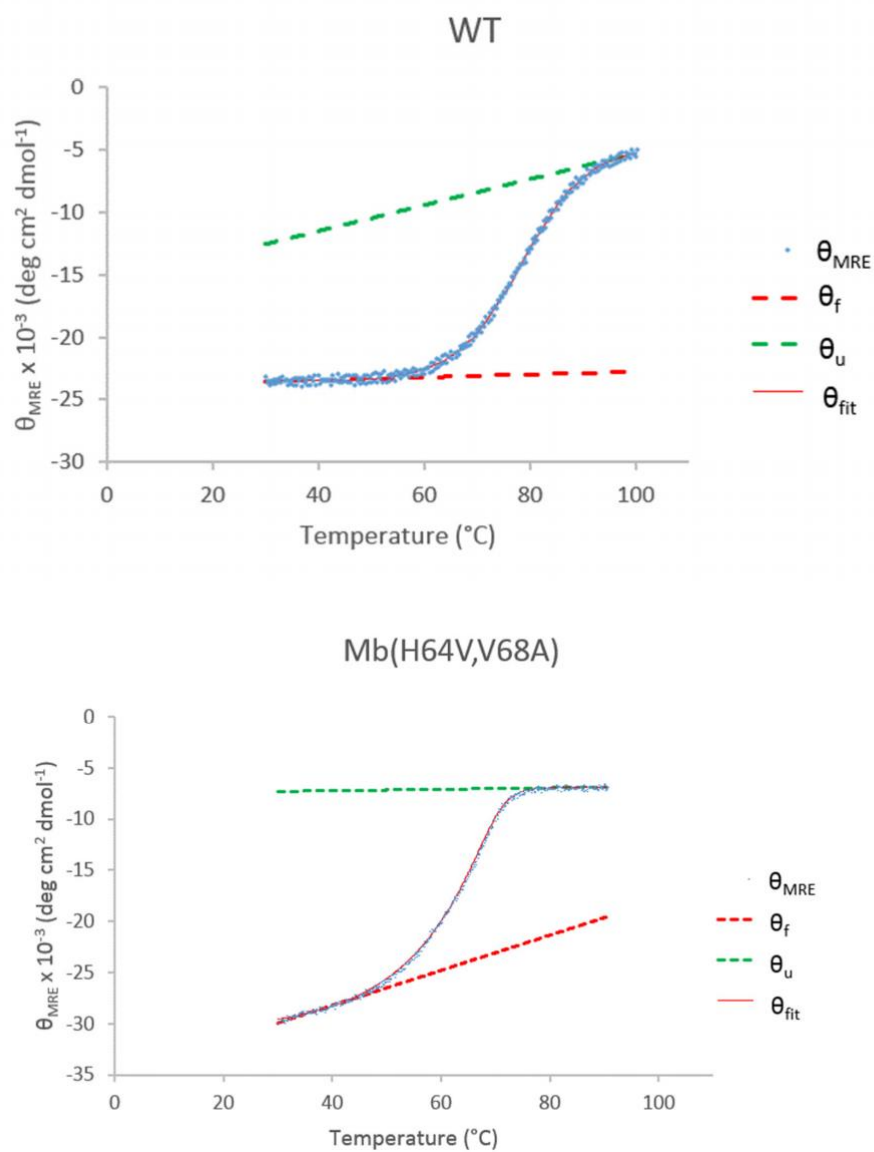


Figure 22-4. Figure S6 (cont.).



**Figure 23-4.** Figure S7. Thermal denaturation curves for wild-type Mb, Mb(H64V,V68A) and sMb variants.

For each variant, a single set of raw data ( $\theta_{MRE}$ ) is shown along with extrapolated signals for folded ( $\theta_f$ ) and unfolded ( $\theta_u$ ) protein and the fitting curve ( $\theta_{fit}$ ).

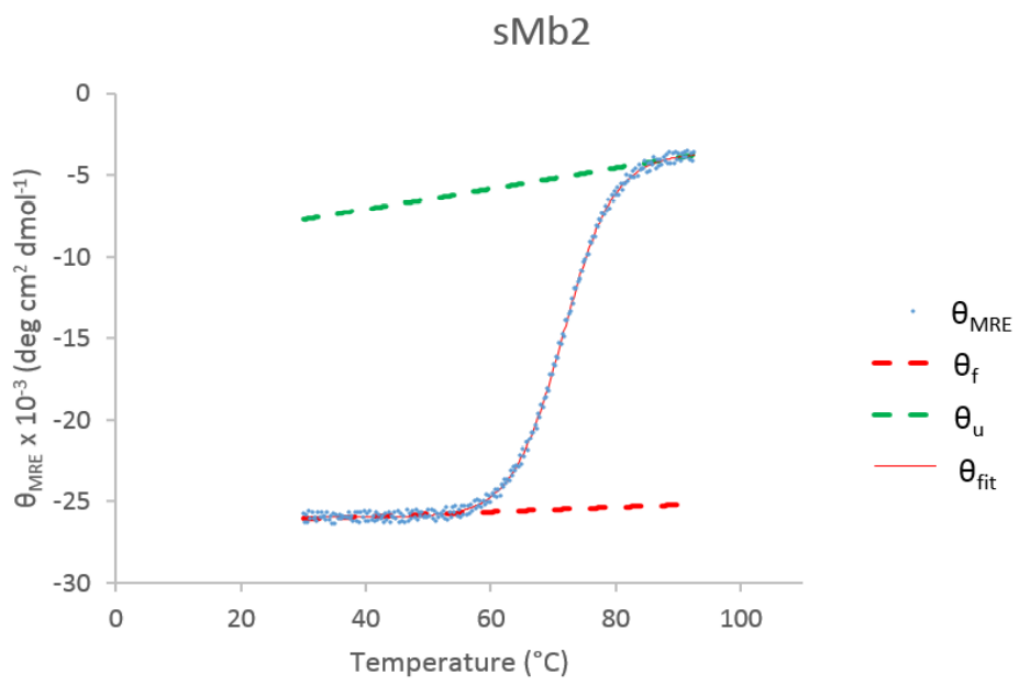
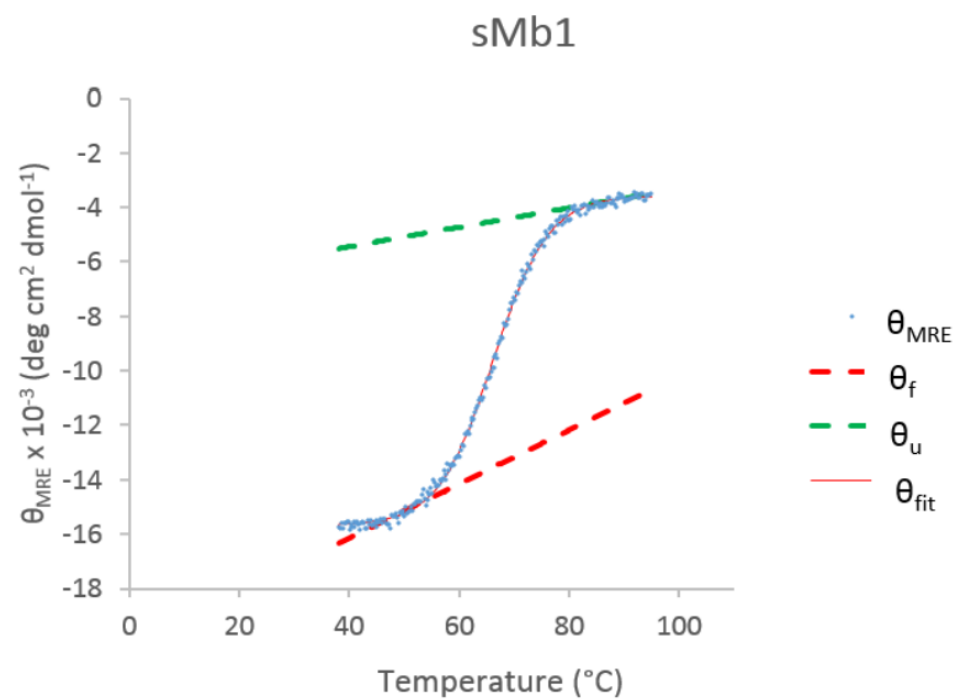


Figure 24-4. Figure S7 (cont.).

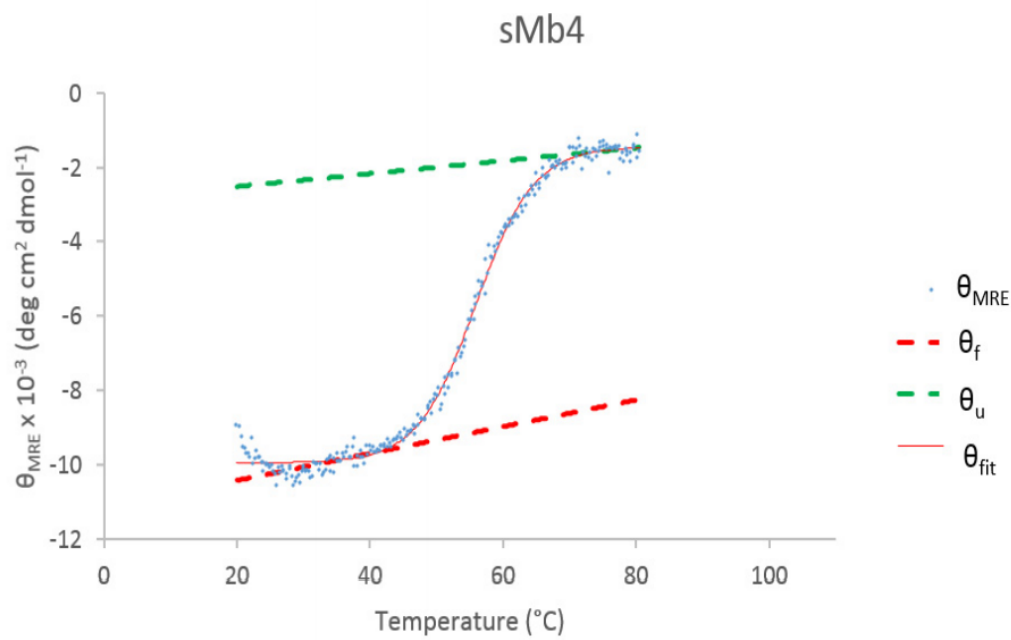
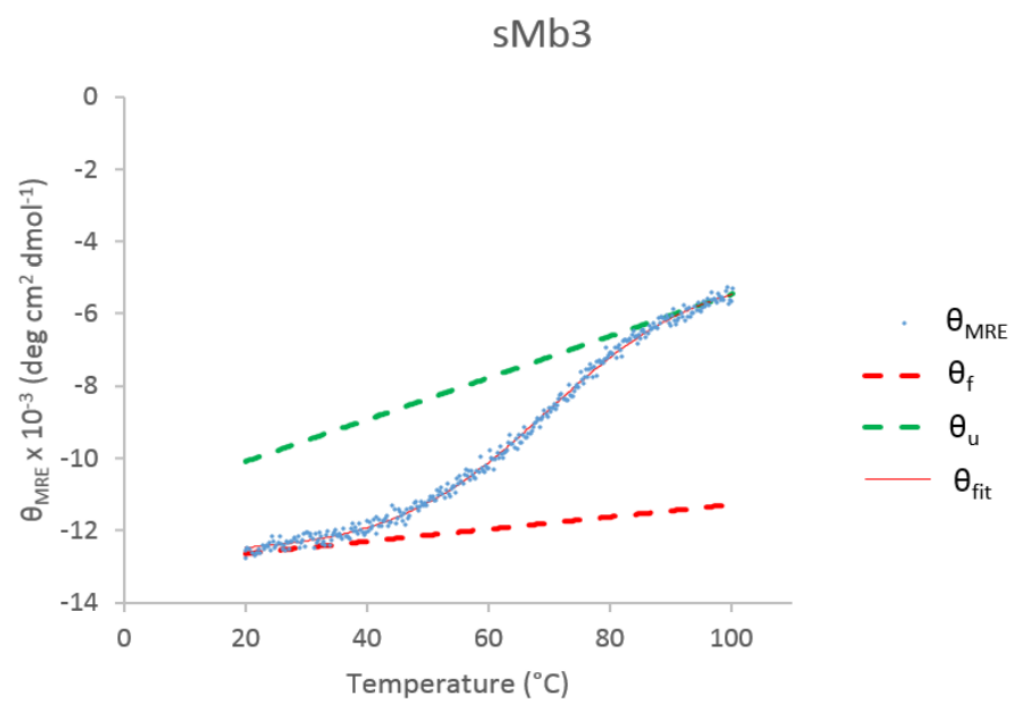


Figure 25-4. Figure S7 (cont.).

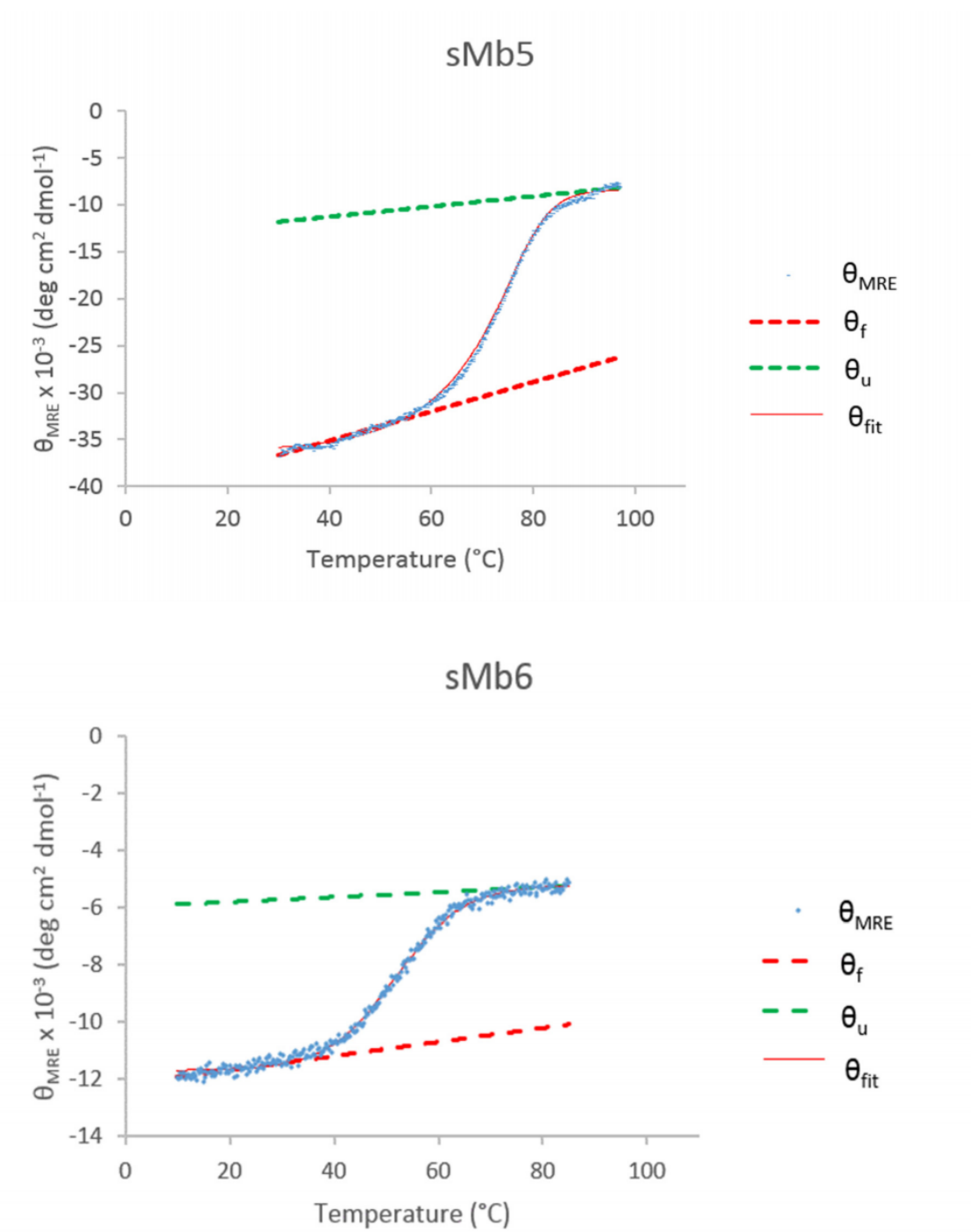


Figure 26-4. Figure S7 (cont.).

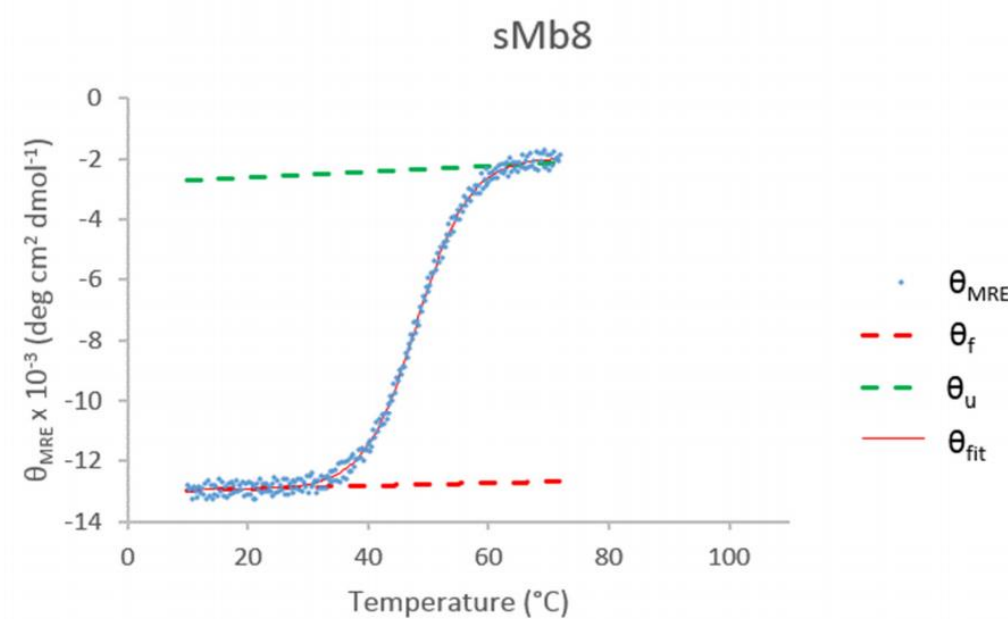
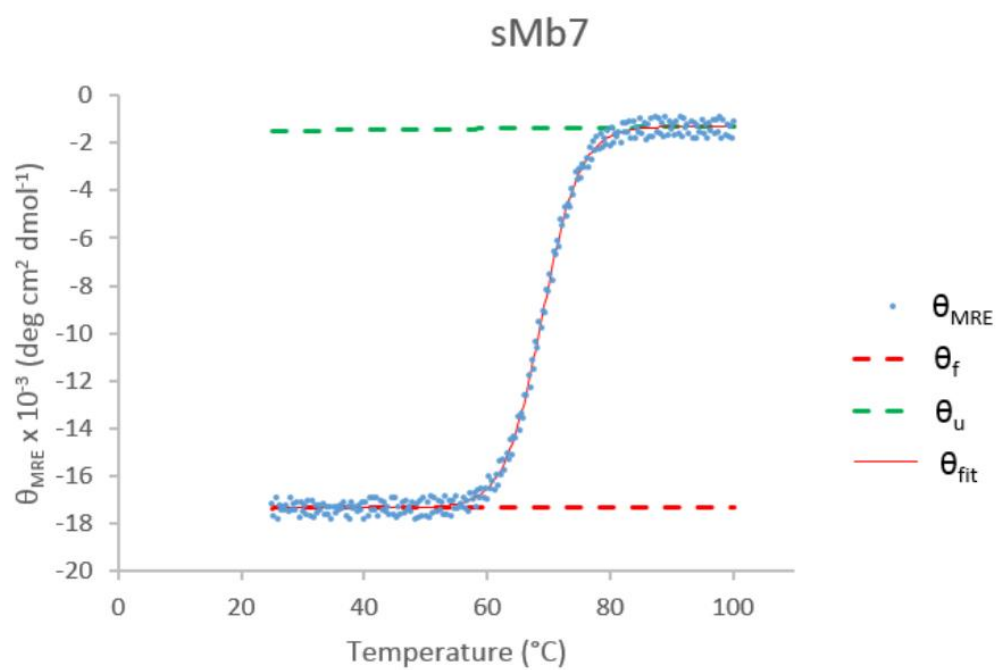


Figure 27-4. Figure S7 (cont.).

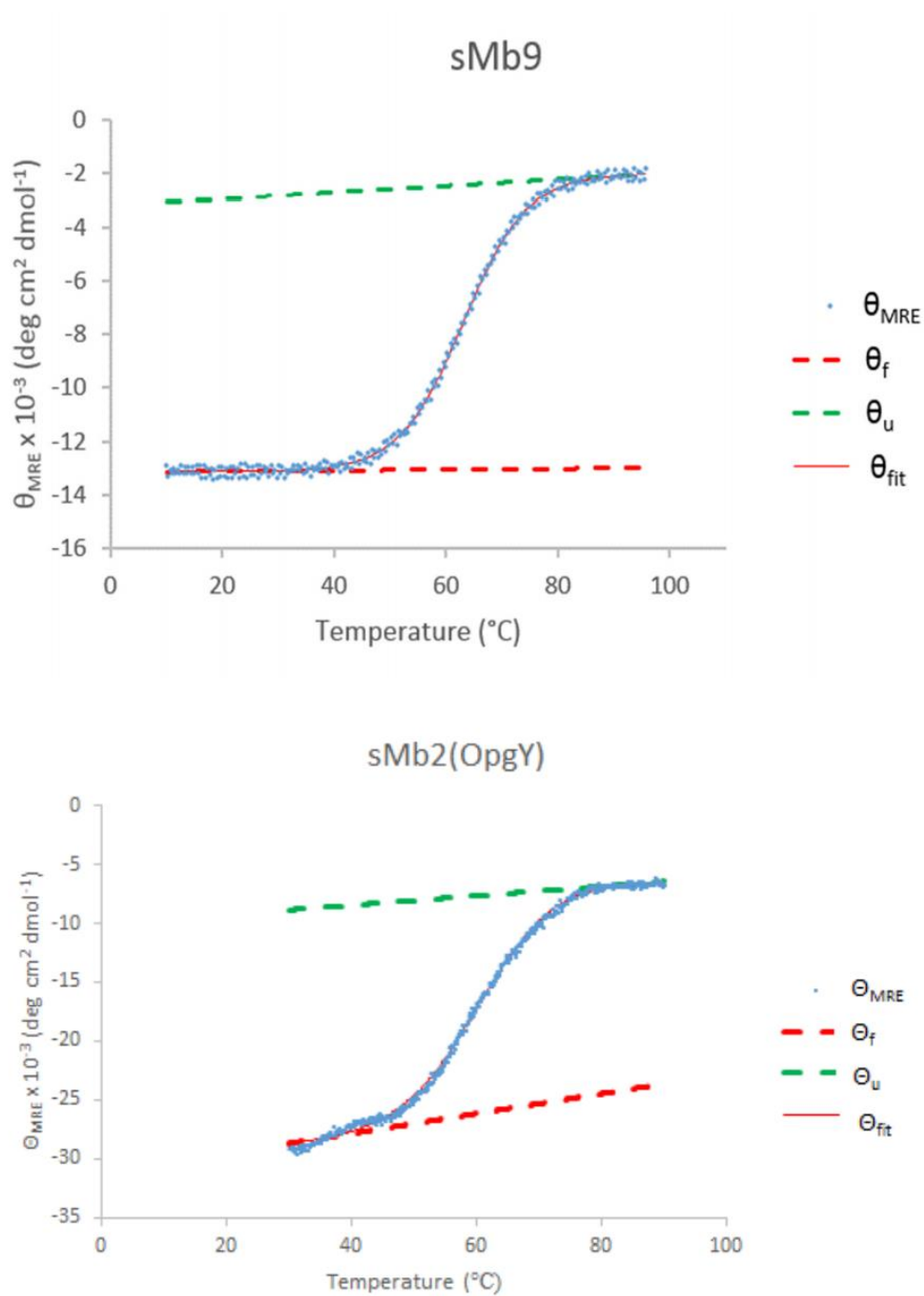


Figure 28-4. Figure S7 (cont.).

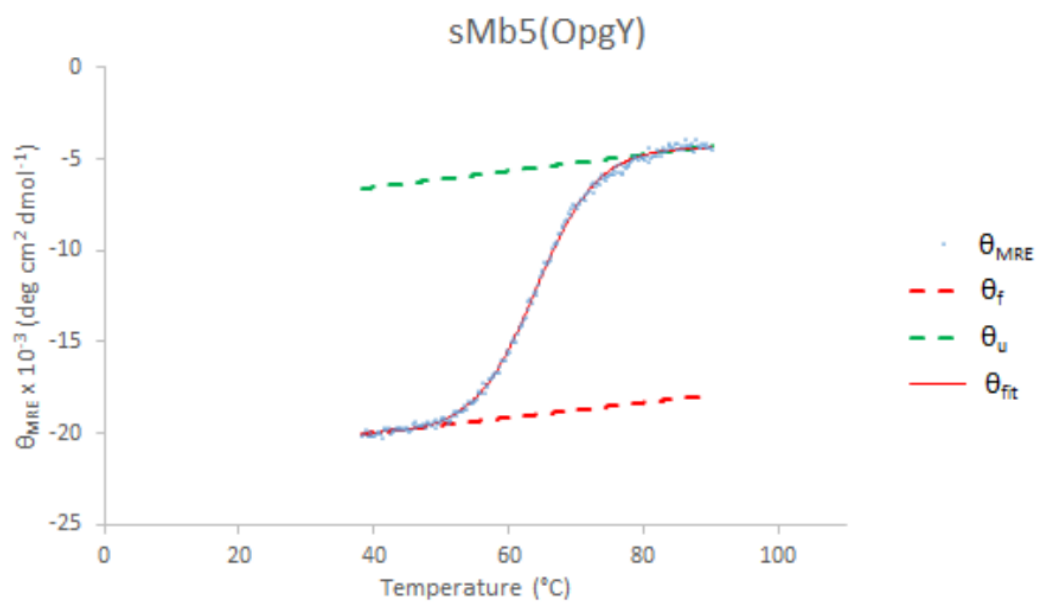


Figure 29-4. Figure S7 (cont.).

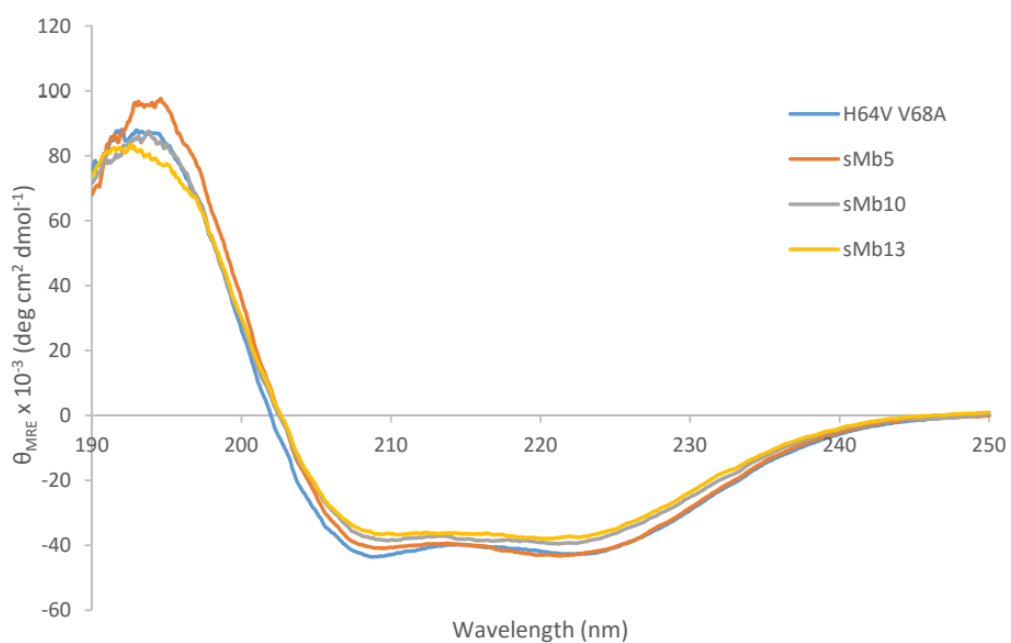
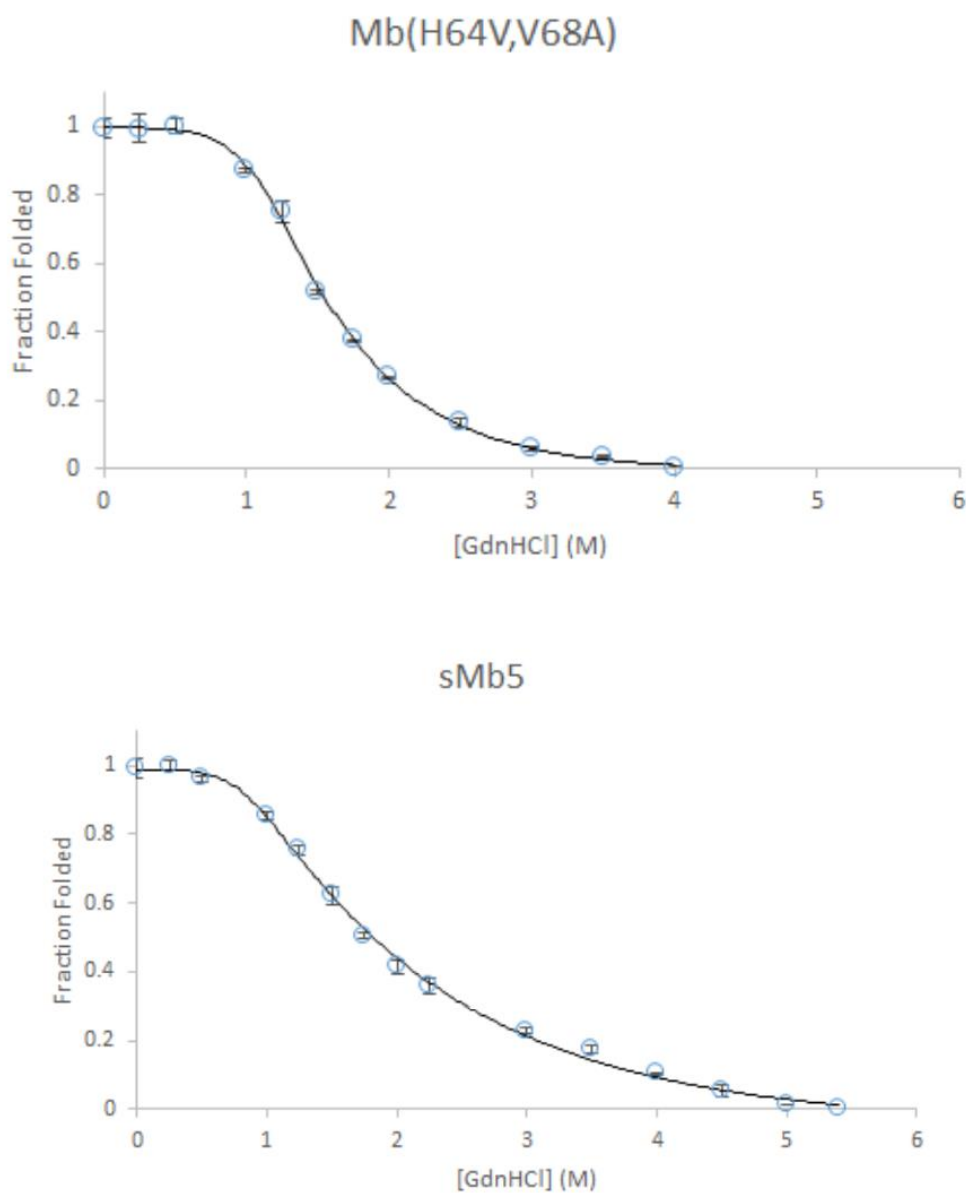


Figure 30-4. Figure S8. Overlay of near-UV circular dichroism spectra corresponding to Mb(H64V,V68A) and stapled variants sMb5, sMb10, and sMb13.



**Figure 31-4.** Figure S9. Chemical denaturation curves for Mb(H64V,V68A) and stapled variants sMb5, sMb10, and sMb13 in the presence of guanidium chloride.

Fraction folded was determined by CD based on change of ellipticity at 220 nm ( $\theta_{220}$ ) at increasing concentrations of denaturant.

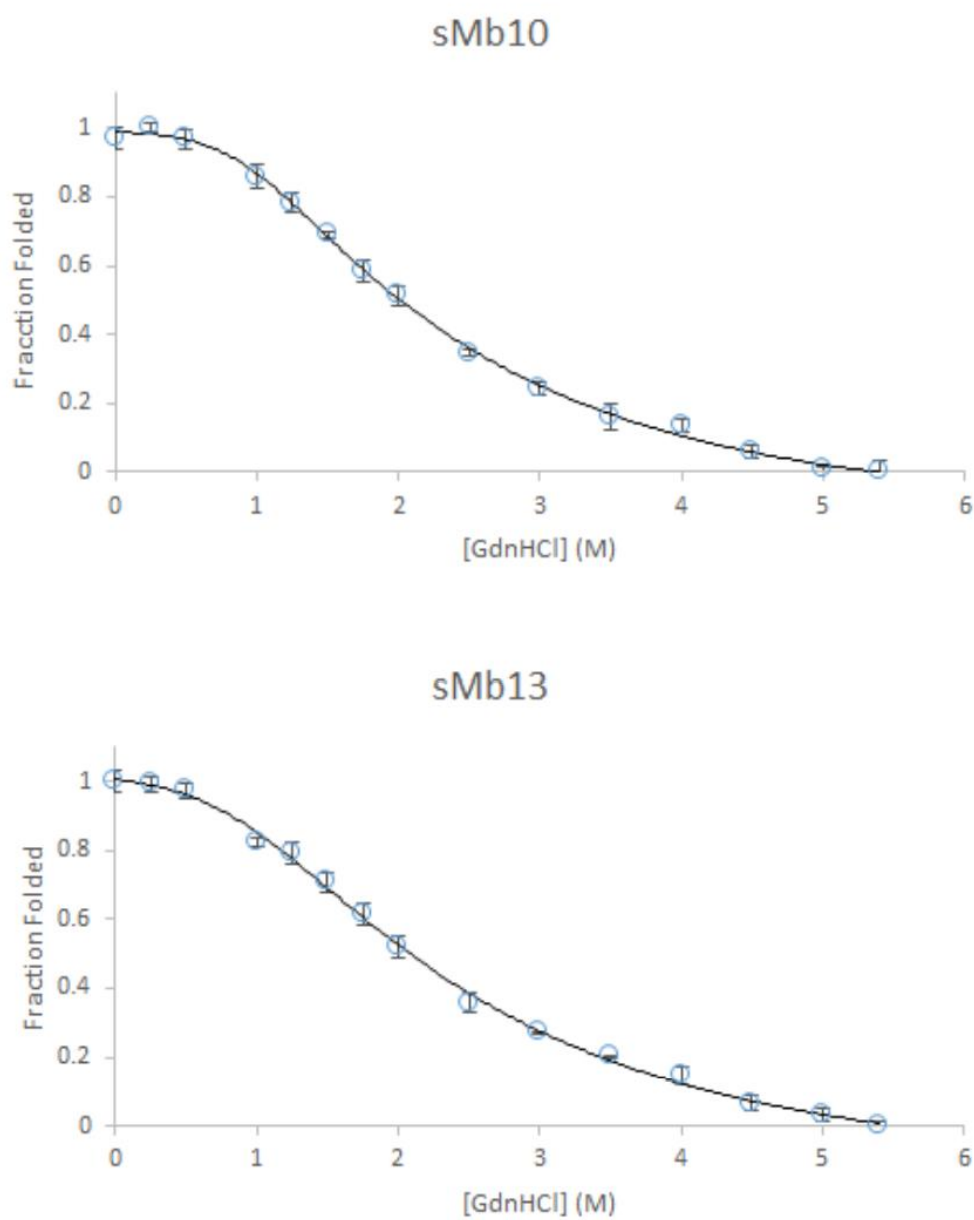


Figure 32-4. Figure S9 (cont.).

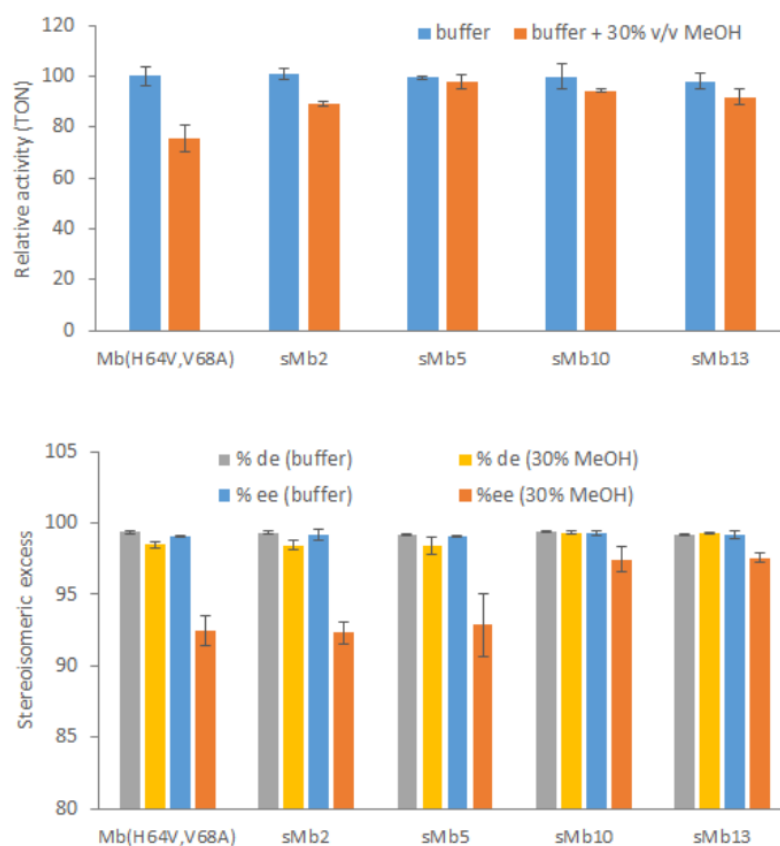
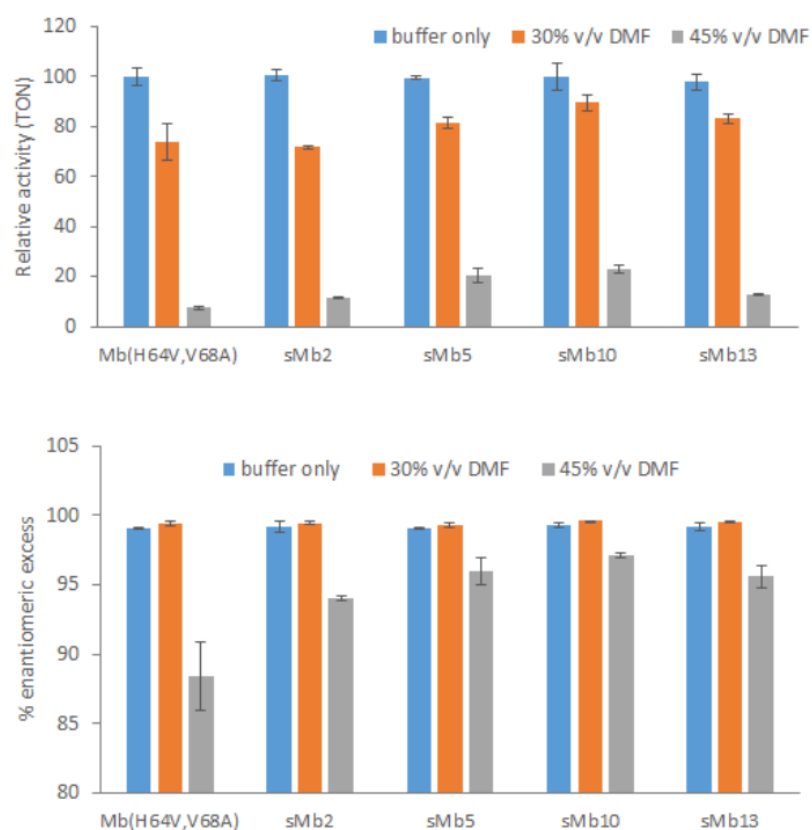


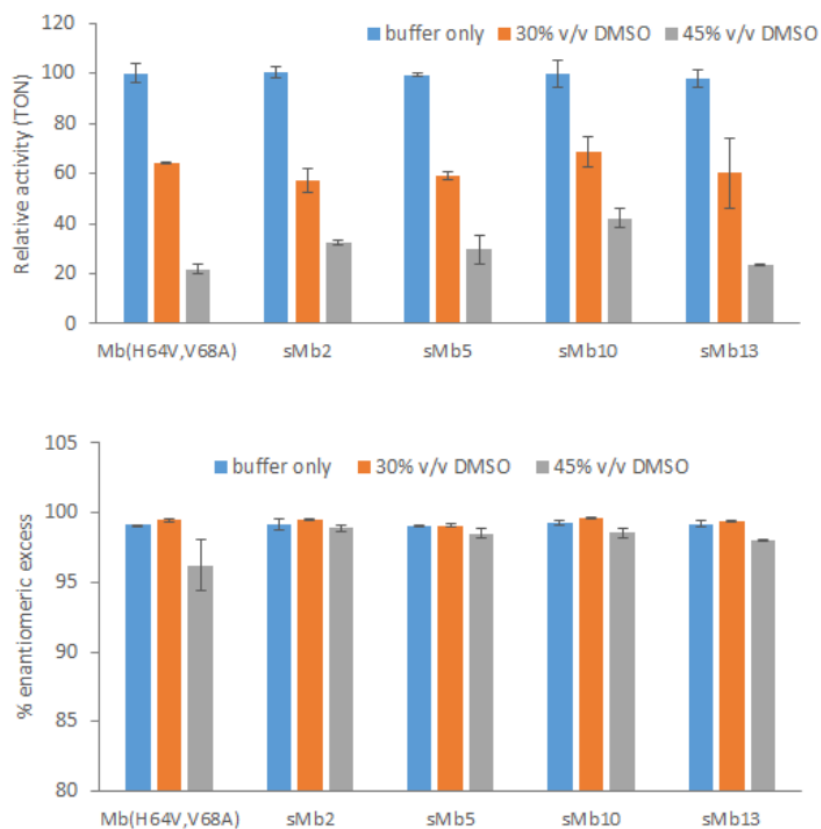
Figure 33-4. Figure S10. Catalytic activity (top graph) and diastereo- and enantioselectivity (bottom graph) of Mb(H64V,V68A) and stapled variants in styrene cyclopropanation reaction in the presence of 30% v/v methanol (MeOH).

Relative activities refer to catalytic turnovers (TON) normalized to TON measured with Mb(H64V,V68A) in buffer only reactions. These data show the higher activity and enantioselectivity of sMb10 and sMb13 compared to the parent protein in the presence of methanol.



**Figure 34-4.** Figure S11. Catalytic activity (top graph) and enantioselectivity (bottom graph) of Mb(H64V,V68A) and stapled variants in styrene cyclopropanation reactions in the presence of 30% and 45% v/v DMF.

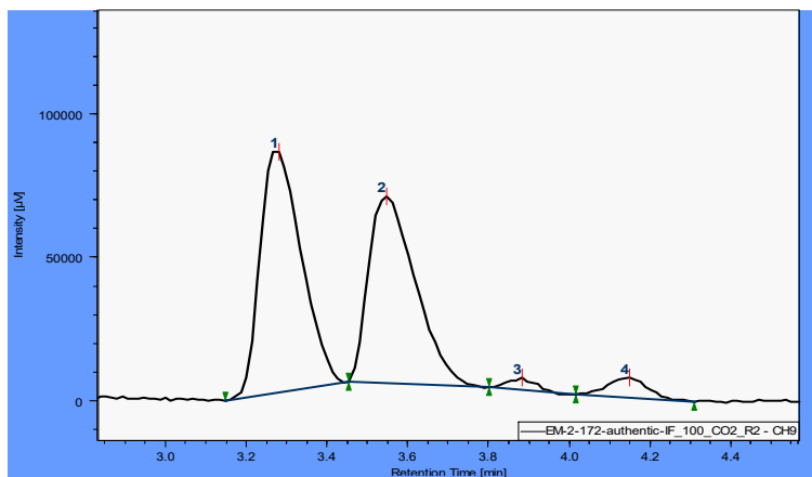
Relative activities refer to catalytic turnovers (TON) normalized to TON measured with Mb(H64V,V68A) in buffer only reactions. These data show the higher activity and enantioselectivity of sMb10 and sMb13 compared to the parent protein in the presence of DMF. Diastereomeric excess (de) values are > 99.5% in all cases and are not shown.



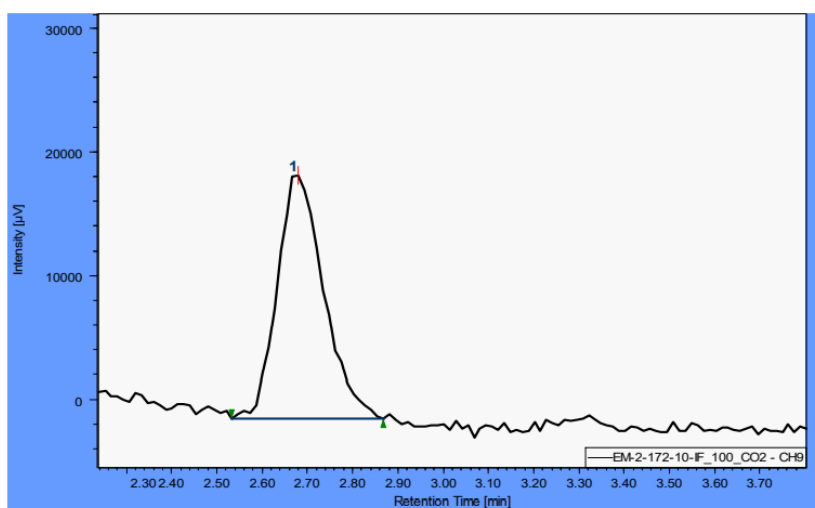
**Figure 35-4.** Figure S12. Catalytic activity (top graph) and enantioselectivity (bottom graph) of Mb(H64V,V68A) and stapled variants in styrene cyclopropanation reactions in the presence of 30% and 45% v/v DMSO.

Relative activities refer to catalytic turnovers (TON) normalized to TON measured with Mb(H64V,V68A) in buffer only reactions. These data show the higher activity and enantioselectivity of sMb10 and sMb13 compared to the parent protein in the presence of DMF. Diastereomeric excess (de) values are > 99.5% in all cases and are not shown.

A)



B)



**Figure 36-4.** Figure S13. SFC analysis of cyclopropanation products from reaction with pentafluorostyrene and EDA.

A) Racemic ethyl 2-(perfluorophenyl)cyclopropane-1- carboxylate (3b) prepared synthetically (trans:cis mixture). B)

Enantioenriched (1S,2S)- 3a produced by reaction with sMb10 (see Table 2 for details on reaction conditions).

Table 7-4. Table S1. Oligonucleotide sequences.

Primer	Sequence (5' to 3')
Myo_XbaI for	TTCCCCTCTAGAAATAATTTTGTTTAAC
Myo_XhoI rev	TTAGAGGCCCAAGGGGTAT
Myo_sMb1_R31TAG for	GACATCCTGATCTAGCTGTTCAAATCT
Myo_sMb1_R31TAG rev	AGATTTGAACAGCTAGATCAGGATGTC
Myo_sMb1_S35K for	TCTGTTCAAAAAACACCCGGAA
Myo_sMb1_S35K rev	TTCCGGGTGTTTTTTGAACAGA
Myo_sMb1_sMb2_E109C for	G TTCATCTCTTGCGCTATCATCC
Myo_sMb1_sMb2_E109C rev	GGATGATAGCGCAAGAGATGAAC
Myo_sMb2_H36TAG for	TGTTCAAATCTTAGCCGGAAACC
Myo_sMb2_H36TAG rev	GGTTTCCGGCTAAGATTTGAACA
Myo_sMb3_L9R_H12C for	AATGGCAGCGTGTTCTGTGCGTTTGGGC
Myo_sMb3_L9R_H12C rev	GCCCAAACGCACAGAACACGCTGCCATT
Myo_sMb3_D122T_A127TAG_f or	ATCCGGGTACCTTCGGTGCTGACTAGCAGGG TGC
Myo_sMb3_D122T_A127TAG_r ev	GCACCCTGCTAGTCAGCACCGAAGGTACCC GGAT
Myo_sMb4_D27TAG for	TCACGGTCAGTAGATCCTGATCC
Myo_sMb4_D27TAG rev	GGATCAGGATCTACTGACCGTGA
Myo_sMb4_H113C_V114G for	GCTATCATCTGCGGTCTGCACT
Myo_sMb4_H113C_V114G rev	AGTGCAGACCGCAGATGATAGC
Myo_sMb5_G5TAG for	TCTGTCTGAATAGGAATGGCAGC
Myo_sMb5_G5TAG rev	GCTGCCATTTCCTATTCAGACAGA
Myo_sMb5_D126C for	CGGTGCTTGCGCTCAGG
Myo_sMb5_D126C rev	CCTGAGCGCAAGCACCG
Myo_sMb6_G25C_I28A for	TGCTGGTCACTGTCAGGACGCC
Myo_sMb6_G25C_I28A rev	GGCGTCCTGACAGTGACCAGCA
Myo_sMb6_L69A for	GTTACCGCGGCGACCG
Myo_sMb6_L69A rev	CGGTGCGCCGCGGTAAC
Myo_sMb6_I111TAG for	TCTCTGAAGCTTAGATCCACGTTCT
Myo_sMb6_I111TAG rev	AGAACGTGGATCTAAGCTTCAGAGA
Myo_sMb7_D20S_G23C_D27A_ for	TTGAAGCTAGCGTTGCTTGTCACGGTCAGGC GATCCTGAT
Myo_sMb7_D20S_G23C_D27A_ rev	ATCAGGATCGCCTGACCGTGACAAGCAACG CTAGCTTCAA
Myo_sMb7_R118TAG for	TCTGCACTCTTAGCATCCGGGT
Myo_sMb7_R118TAG rev	ACCCGGATGCTAAGAGTGCAGA
Myo_sMb8_H24TAG_I28S for	TGCTGGTTAGGGTCAGGACAGCCTGATC
Myo_sMb8_H24TAG_I28S rev	GATCAGGCTGTCCTGACCCTAACCAGCA
Myo_sMb8_L69S for	TTACCGCGAGCACCGCTC
Myo_sMb8_L69S rev	GAGCGGTGCTCGCGGTAA

Myo_sMb8_I111C_for	CTCTGAAGCTTGCATCCACGTT
Myo_sMb8_I111C_rev	AACGTGGATGCAAGCTTCAGAG
Myo_sMb9_K16C_for	TTTGGGCTTGCGTTGAAGCT
Myo_sMb9_K16C_rev	AGCTTCAACGCAAGCCCCAAA
Myo_sMb9_H119A_G121TAG_D122S_for	CTCTCGTGCGCCGTAGAGCTTCGGTGC
Myo_sMb9_H119A_G121TAG_D122S_rev	GCACCGAAGCTCTACGGCGCACGAGAG
Myo_sMb11_F106A_for	AATACCTGGAGGCCATCTCTTGC
Myo_sMb11_F106A_rev	GCAAGAGATGGCCTCCAGGTATT
Myo_sMb12_G129E_for	TTGCGCTCAGGAAGCTATGAACA
Myo_sMb12_G129E_rev	TGTTCATAGCTTCCTGAGCGCAA
Myo_sMb13_H113E_for	GCGCTATCATCGAAGTTCTGCACTC
Myo_sMb13_H113E_rev	GAGTGCAGAACTTCGATGATAGCGC

**Table 8-4.** Table S2. Rosetta scores of designs sMb1-9 after energy minimization and rotameric sampling.

Scores are shown in Rosetta Energy Units (R.e.u.) (a combined score of Rosetta physical and statistical modeling). Total scores refer to the residues CYZ (attacking deprotonated cysteine), TYZ (backbone attached portion of O2beY, corresponding to deprotonated tyrosine residue), and XLB (thioether bridge portion of O2beY/Cys crosslink). The geometric constraints score is also shown. Higher residue scores represent greater deviation from expected residue conformation or clashes; higher constraint score represents greater deviation from geometric constraints.

Variant	Mutations	CYZ Rosetta residue score (R.e.u.)	TYZ Rosetta residue score (R.e.u.)	XLB Rosetta residue score (R.e.u.)	Geometric constraint score (R.e.u.)
sMb1	R31(O2beY), S35K, E109C	1.4	-0.1	0.3	3.1
sMb2	H36(O2beY), E109C	-1.7	0.0	0.1	0.9
sMb3	L9R, H12C, D122T, A127(O2beY)	0.1	0.2	4.5	4.0
sMb4	D27(O2beY), H113C, V114G	2.7	1.3	3.6	4.0
sMb5	G5(O2beY), D126C	0.8	-0.3	0.2	0.9
sMb6	G25C, I28A, L69A, I111(O2beY)	14.7	1.7	19.2	4.4
sMb7	D20S, G23C, D27A, R118(O2beY)	0.0	1.2	2.1	1.9
sMb8	H24(O2beY), I28S, L69S, I111C	14.7	1.7	19	4.4
sMb9	K16C, H119A, G121(O2beY), D122S	8.2	1.2	65.6	10.8

## 4.8 Supplemental References

1. Kondo H, Nakamura T, & Tanaka S: **A significant role of Arg41 residue in the enzymatic reaction of haloacid dehalogenase L-DEX YL studied by QM/MM method.** *J Mol Catal B-Enzym* (2014), **110**:23-31.
2. Bhattacharya S & Lecomte J: **Temperature dependence of histidine ionization constants in myoglobin.** *Biophys. J.* (1997), **73**(6):3241-3256.

3. Zanghellini A, et al.: **New algorithms and an in silico benchmark for computational enzyme design.** *Protein Sci* (2006), **15**(12):2785-2794.
4. Richter F, Leaver-Fay A, Khare SD, Bjelic S, & Baker D: **De novo enzyme design using Rosetta3.** *PloS one* (2011), **6**(5):e19230.
5. Bordeaux M, Tyagi V, & Fasan R: **Highly Diastereoselective and Enantioselective Olefin Cyclopropanation Using Engineered Myoglobin-Based Catalysts.** *Angew Chem Int Ed Engl* (2015), **54**(6):1744-1748.
6. Bionda N, Cryan AL, & Fasan R: **Bioinspired strategy for the ribosomal synthesis of thioether-bridged macrocyclic peptides in bacteria.** *ACS Chem Biol* (2014), **9**(9):2008-2013.

## **Chapter 5:      A site-selective protein modification for N<sub>2</sub>S\* metal chelation**

### **5.1 Preface**

This chapter is being prepared for publication.

### **5.2 Abstract**

Precise metal-protein coordination by design remains a considerable challenge. High-affinity polydentate protein modifications, chemical and recombinant, enhance metal localization but are often bulky, flexible, or coordinately saturated. Here, we expand the biomolecular toolbox with a simple two step synthesis of bis(1-methyl-1H-imidazole-2-yl)ethene or BMIE, a compact high-affinity metal-coordination ligand capable of thiol conjugation. The conjugation of BMIE to various small-molecule thiols (l-cys, boc-cys, thiocresol, and glutathione) confirm general thiol reactivity. The small-molecule thiol adducts are shown to coordinate multiple divalent metal ions (Co, Ni, Cu, and Zn) in bidentate (N<sub>2</sub>) and tridentate (N<sub>2</sub>S\*) coordination geometries. Site-selective BMIE modification (>90% yield at pH 8.0) of a carboxypeptidase G2 variant (S203C), measured with ESI-MS, confirms utility as a bioconjugate. EPR characterization of modified protein and CuCl<sub>2</sub> reveal site selective 1:1 BMIE-Cu<sup>2+</sup> coordination and symmetric tetragonal geometry under physiological conditions with a host of counter-ligands (H<sub>2</sub>O/HO<sup>-</sup>, tris, and phenanthroline). Furthermore, x-ray derived structures of modified protein co-crystallized with ZnCl<sub>2</sub> reveal a heterogeneous BMIE-Zn<sup>2+</sup> contact along the intermolecular crystallographic interface between a modified and an unmodified cysteine monomer. These features, combined with ease of synthesis, make

for an attractive metalloprotein design tool and should enable future catalytic and structural applications.

### 5.3 Introduction

Accurate metalloprotein design remains a considerable challenge. Often, the intended coordination geometry differs from the observed<sup>1</sup>. The successful design of metalloproteins requires the design of multiple coordinating amino acids with exacting coordination constraints, first shell hydrogen bonding interactions to position the coordinating amino acids, and a hydrophobic second shell to promote polarizability and metal affinity<sup>2,3</sup>. To reduce the complexity of design and promote metal binding, successful strategies utilize polydentate artificial protein modifications such as cofactors, site selective chemical modification, or dative unnatural amino acids (UAAs)<sup>4,5,6</sup>.

Site-selective chemical modifications offer an attractive way to functionalize a protein for desired metal-binding. However, compared to UAAs, their bulky size and flexible nature complicate protein design, which increases the need for compensatory mutations<sup>7,8,9</sup>. As an example, a commonly used polydentate cysteine modification, Iodoacetemido-1,10-phenanthroline (iodo-phen), can replace two natural amino acids but contains 20 non-backbone heavy atoms and 5 torsional degrees of freedom when conjugated to cysteine. Alternatively, UAAs simplify design but suffers from reduced protein yields and undesirable endogenous metal incorporation<sup>5,6</sup>.

To facilitate metalloprotein design, we sought to combine the benefits of the different artificial protein modification strategies to create a compact cysteine-selective conjugate

addition that would enable high-affinity metal coordination. We chose to design a bidentate n-heterocyclic amine functional group, similar to bpy-ala and other reported successful protein modifications. Bidentate n-heterocyclic amines offer high metal affinity and a diverse metal-ion binding pallet. Methylation of the imidazole groups further increases metal affinity. Cysteine was selected as the target residue for labeling due to the relatively low abundance in natural proteins, selective nucleophilicity at physiological pH, and small contributions to both steric bulk and eventual flexibility. Commonly used cysteine anchoring strategies include halo-alkyl (iodoacetamide) and  $\alpha,\beta$ -unsaturated systems (vinyl sulfones and malaemides)<sup>10</sup>. To reduce steric size and number of rotatable bonds, we encoded the anchoring functional group directly into the bidentate n-heterocyclic amine by using an  $\alpha,\beta$ -unsaturated vinylidene as an anchor. Five membered n-heterocyclic amines, such as imidazole, are significantly more electrophilic at physiological conditions than six membered n-heterocycles such as pyridine. In addition, introducing a bridging vinylidene would allow for shared coordination of a proton, between the n-heterocyclic amines further increasing the pKa and electrophilicity under physiological conditions, suitable for conjugation with a nucleophilic cysteine.

## 5.4 Results and Discussion

### 5.4.1 Preparation of small-molecule BMIE-thiol adducts

To explore general thiol-reactivity in the presence of BMIE, we prepared several small-molecule BMIE-thiol adducts: BMIE-thiocresol<sub>MeOH</sub> (**2a**) and BMIE-boc-cys<sub>MeOH</sub> (**2b**) and BMIE-l-cys<sub>aq</sub>. Adducts **2a** and **2b** were prepared in methanol at ambient temperature for 2 hours, following the scheme in Figure 1a and confirmed using MS and <sup>1</sup>H/<sup>13</sup>C NMR

(SF1 & 2). Conjugation of BMIE with l-cys was performed in aqueous conditions (0.1mM NaPO<sub>4</sub>, 1mM EDTA, pH8) at ambient temperature. The extent of reaction with l-cys was measured over time with Ellman's reagent and proceeded to completion within 4 hours (SF3). Optimization of the adduct formation conditions are further described in the methods section of the supplemental material.

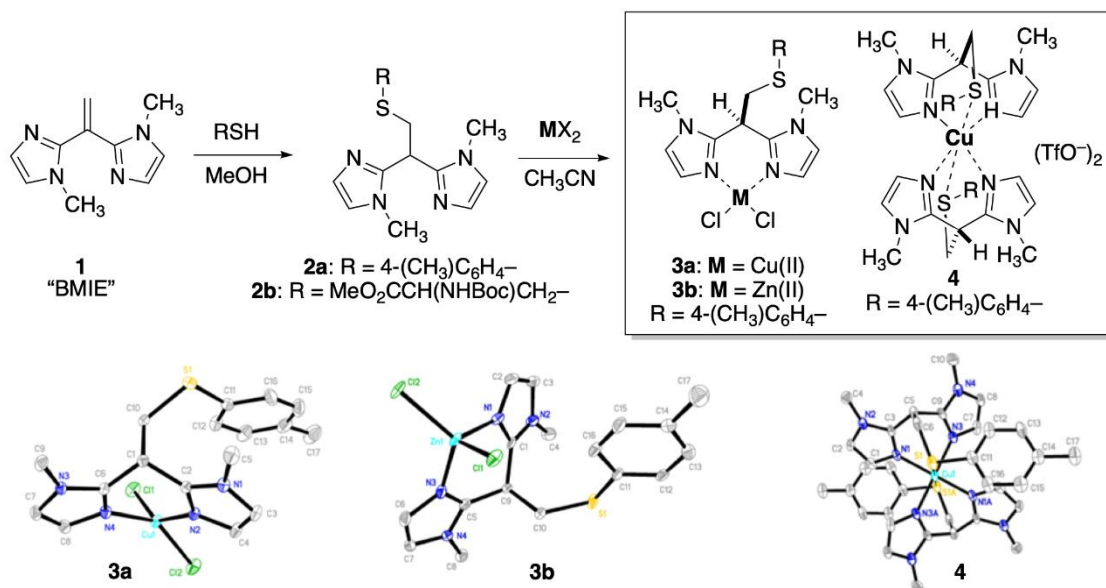
#### 5.4.2 Small-molecule BMIE-thiol metal coordination

*UV/vis.*— Following preparation, BMIE-cys was incubated with several divalent transition metals (Co, Ni, Cu, and Zn) and measured using UV/vis (SF4). The appearance of MLCT peaks in the visible region for M<sup>2+</sup>:BMIE-cys suggests, by first approximation, qualitative metal coordination by BMIE-cys and corroborates the robust metal coordination observed by Varnagy et al. Concentrations in excess of 2mM were necessary to meet the detection limit of our instrument. Although we did not determine the exact MLCT extinction coefficient, we find our results for Cu<sup>2+</sup>:BMIE-cys agree with extinction coefficients reported in the literature for similar complexes<sup>11</sup>. Unfortunately, protein concentrations seldom exceed 0.5mM and remain stable. Therefore, alternative methods may be necessary to characterize BMIE-protein metal coordination. Furthermore, Zn<sup>2+</sup> will not participate in MLCT and should also be analyzed with alternative methods.

*NMR.*— To characterize Zn<sup>2+</sup> and Cu<sup>1+</sup> coordination and stoichiometry, titrations (0 to 2.5 molar equivalents) of ZnCl<sub>2</sub>, Zn(II) triflate, and Cu(I) tetrafluoroborate in the presence of **2b** were analyzed with <sup>1</sup>H NMR. Downfield aliphatic and aromatic chemical shifts were observed across all metal salt titrations from 0 to 1 molar equivalents of metal salt (SF5-

7). Further addition of metal salt resulted only in peak sharpening, suggesting 1:1 complexation with **2b** and excess metal. These results highlight BMIE as an alternative to other protein modification analogs (i.e., BPY derivatives) ill-suited for coordinating  $\text{Zn}^{2+}$  and  $\text{Cu}^{1+}$ . We suspect the wider "bite"-angle, combined with the softer 5-membered heterocyclic amines, for the improved coordination.

*X-ray crystallography.*— Following 1:1 mixing of **2a** with several salts ( list all), evaporation led to three unique structures (Figure 1b).  $\text{ZnCl}_2$  and  $\text{CuCl}_2$  salts produced four-coordinate structures with 1:1 adduct/metal bidentate ( $\text{N}_2$ ) coordination and two bound Cl counterions. While  $\text{Zn}(\mathbf{2a})(\text{Cl})_2$  coordination adopts a typical tetrahedral geometry,  $\text{Cu}(\mathbf{2a})(\text{Cl})_2$  exhibits an oblique  $45^\circ$  distortion of the Cl-Cu-Cl plane (SF8). We suspect the observed distortion is the result of Cu(II) preference for square planar/pyramidal coordination geometries, the hardness of the chloride counterion, and steric strain to form a plane with present counterions. Crystals of **2a** and Cu(II) triflate formed 2:1 adduct/Cu complexes,  $\text{Cu}(\mathbf{2a})_2$ , with six-coordinate tetragonal geometry. Each ligand formed a tridentate ( $\text{N}_2\text{S}^*$ ) coordination with four coordinating nitrogen atoms in the equatorial plane and two thioether sulfur at each axial position. Three isoforms of the tetragonal geometry were solved with varying Cu-S distances ranging from 2.94-3.16 Å. These results suggest that the metal-coordination environment is strongly influenced by the counterions present. Furthermore, BMIE adducts can participate in tetrahedral and tetragonal coordination geometries.



**Figure 37-5.** Figure 1. BMIE characterization with small molecule BMIE-thiol adducts.

(A) Time dependent conjugation of BMIE to l-cysteine measured using uv-vis spectroscopy. The % labeling of small molecules was calculated from the left over available free-thiol in solution measured by the absorbance at 412 after adding Ellman's reagent. (B) Uv-vis spectra of BMIE-l-cys:M<sup>2+</sup> compounds after 5 minute incubation at RT. (C) NMR – multi-trace Zn and Cu titration experiment (D) BMIE-TC:Zn<sup>2+</sup> crystal structure formed from BMIE-TC and ZnCl<sub>2</sub>. (E) BMIE-TC:Cu<sup>2+</sup> crystal structure formed from BMIE-TC and CuCl<sub>2</sub>. (F) Three isoforms of (BMIE-TC)<sub>2</sub>:Cu<sup>2+</sup> formed from BMIE-TC and Cu-triflate.

#### 5.4.3 BMIE modification of Carboxypeptidase G2

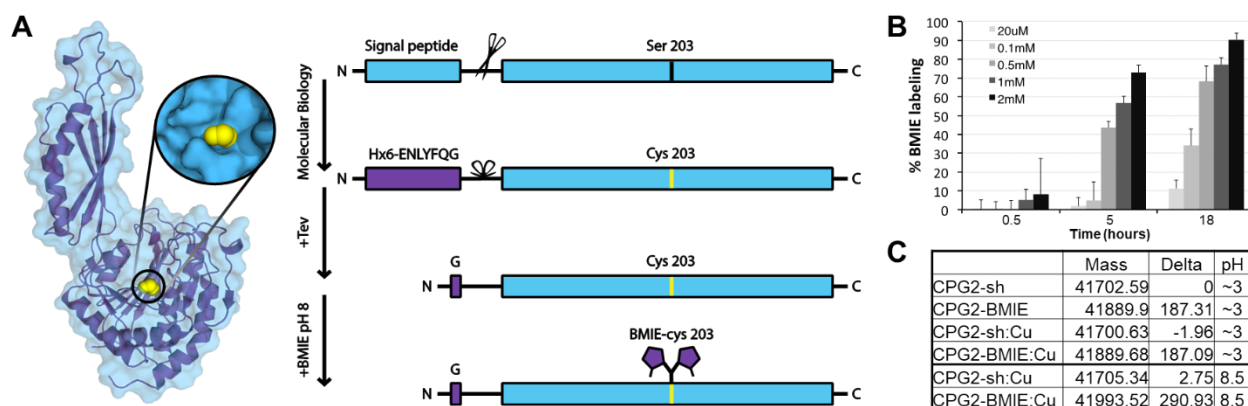
To show that BMIE can be utilized as a protein modification tool, we made a variant of wildtype Carboxypeptidase G2 with an S203C substitution (CPG2-S203C, Figure 2a) and replaced the first 25 amino-acids, known to facilitate periplasmic signaling, with a TEV-cleavable his-tag (Figure 2b). Modification of CPG2-S203C with BMIE was performed in 100mM NaPO<sub>4</sub>, 1mM EDTA, at pH8. The extent of reaction was measured as loss of free cysteine measured over time with Ellman's reagent (fig. 2b). The time-dependent

modification assay revealed a concentration dependent formation of BMIE-CPG2-S203C; the fastest rate (200 molar excess BMIE) proceeded to ~70% modification after 4 hours at room temperature and >90% after 18 hours (fig. 2b and SF9). Similar modification efficiency was observed when samples were first labeled at room temperature for 4 hours followed by labeling overnight at 4C. ESI-MS confirmed a single (+188 Dalton) site-specific modification of CPG2-S203C at position 203 (fig. 2c). Protein purification, modification conditions, and ESI-MS conditions are described in more detail in the methods section within supporting information.

#### 5.4.4 Qualitative metal coordination by BMIE-CPG2-S203C

ESI-MS and ICP-MS were used to qualitatively determine Cu and Zn binding to CPG2-BMIE. His-tags were removed via incubation with tev protease. Metal binding buffers (10mM tris, 100mM NaCl pH 7.4 and 8.5) used to prepare the samples were treated with Chelex-100 resin to remove trace metal ions. After separate incubation with metal salts (5 eqv.  $\text{ZnCl}_2$  and 3 eqv.  $\text{CuCl}_2$ ) for 30 minutes at ambient temperature, the samples were desalted with a PD10 column and eluted with Chelex-100 treated buffer. Direct infusion (detailed in the supplemental methods) of the Cu-incubated modified protein at pH 8.5 revealed a ~290 dalton increase from unmodified protein (Fig 2c). This change in mass supports a modified protein (+188), a bound Cu ion (+65), and two bound  $\text{H}_2\text{O}/\text{OH}^-$  (+36). 1:1 metal/modified-protein coordination as measured by ICP-MS of the pH 8.5 samples confirmed modification-dependent binding of metal ions. At pH 7.4, we observed <100% metal coordination, which implicate pH or vaporization conditions. In support of this finding, standard ESI-MS conditions, which reduce the pH to ~3.0, were

too harsh to detect bound Cu. These results are consistent with the pKa values (6.5-6.9) of structural BMIE analogs reported in the literature<sup>11</sup>.



#### 5.4.5 EPR characterization of $[\text{Cu}(\text{BMIE-CPG2-S203C})]^{2+}$

**Figure 38-5.** Figure 2. CPG2 modification and site specific cysteine labeling with BMIE.

(A) Crystal structure of CPG2 (PDB: 1CG2) with ser203 highlighted in yellow. Modification of CPG2 by replacement of the N-terminal 25 amino acid signal peptide with a his<sub>6</sub>-tev-cleavage tag and a S203C substitution. After purification of the protein, the N-terminal his<sub>6</sub> tag is cleaved with tev protease and the resulting protein is labeled with BMIE. (B) Formation of CPG2-BMIE after incubation with increasing concentrations of BMIE in 100mM NaPO<sub>4</sub>, 1mM EDTA, pH8 at RT. Amount of labeling was calculated from the available free-thiol in solution measured using Ellman's reagent—absorbance at 412nm. (C) Summary of ESI-MS analysis for labeled and unlabeled CPG2 with or without prior incubation with CuCl<sub>2</sub>. ESI-MS performed under standard conditions (pH ~3) and by direct infusion (pH 8.5) are reported. The delta mass is calculated by subtracting CPG2-sh under standard conditions.

To further validate site specific metal binding under physiological conditions, we evaluated the paramagnetic species  $[\text{Cu}(\text{BMIE-CPG2-S203C})]^{2+}$  and  $[\text{Cu}(\text{BMIE-l-cys})]^{2+}$  for <sup>1</sup>H and <sup>14</sup>N hyperfine interactions within 4Å of the metal center by pulsed-EPR methods. Samples were prepared in Chelex-treated buffer (10mM MOPS, 100mM NaCl at pH 7.5). CW, ESEEM, and ENDOR experiments were used to determine tetragonal coordination of Cu<sup>2+</sup> by multiple <sup>14</sup>N atoms in  $[\text{Cu}(\text{BMIE-l-cys})]^{2+}$ . Nine <sup>14</sup>N hyperfine-

split lines were resolved at  $g_{\perp}$  orientation (330 mT) which support the unique cluster speciation with four equivalent  $^{14}\text{N}$  nuclei (two symmetric BMIE-1-cys ligands) coordinated in an equatorial plane. We observed broad peaks at  $g_{\parallel}$  for  $[\text{Cu}(\text{BMIE-CPG2-S203C})]^{2+}$  and suspected mixed  $\text{H}_2\text{O}/\text{OH}^-$  coordination at pH 7.5. At pH 9.65, the peaks at  $g_{\parallel}$  sharpen and the four  $^{14}\text{N}$  atoms can be resolved as 2 identical coordinating imines (proximal) and 2 identical N-CH<sub>3</sub> amines with a distance of  $3.1 \pm 0.1 \text{ \AA}$  (remote) from the metal, suggesting n-methylated imidazole as the coordinating species. Spin harmonics ( $n \cdot \nu_{dq}^+$ ) measured using three-pulse ESEEM (Figure 3b) confirms 2:1 ligand/Cu (4 harmonics) and 1:1 protein/Cu binding (2 harmonics). These results were confirmed with a Davie's ENDOR experiment (Figure 3c). Simulations (supplemental methods) of CW EPR spectra for all species were produced (dashed lines Figure 3a) and collected in Table 1.

We wanted to show BMIE-CPG2-S203C could retain metal coordination in the presence of exogenous metal-coordinating ligands. We prepared a sample of  $[\text{Cu}(\text{BMIE-CPG2-S203C})]^{2+}$  in the standard MOPS buffer at pH with a molar equivalent of phenanthroline added and a sample where 10mM MOPS was replaced with 10mM Tris. In each sample, CW hyperfine lines were resolved. When compared to the sample in MOPS at pH 9.65, shifts in  $g_{\parallel}$  indicate novel speciation. Similarly, ESEEM and ENDOR experiments confirm BMIE-CPG2-S203C symmetric  $^{14}\text{N}$  coordination at the equatorial positions which suggest the changes in spectra are a product of counter-ligand coordination at the remaining equatorial positions. This is confirmed with an ENDOR comparison showing that  $[\text{Cu}(\text{BMIE-CPG2-S203})(\text{phenanthroline})]^{2+}$  is a 1:1 sum of  $[\text{Cu}(\text{phenanthroline})_2]^{2+}$

and  $[\text{Cu}(\text{BMIE-1-cys})_2]^{2+}$ . Taken together, BMIE-CPG2-S203C has a single and unique tetragonal BMIE-metal binding site that can be tuned with various counter ligands.

#### 5.4.6 X-ray crystallography of $[\text{Zn}(\text{BMIE-CPG2-S203C})]^{2+}$

We solved the crystal structure of  $[\text{Zn}(\text{BMIE-CPG2-S203C})]^{2+}$  at a resolution of 3.1 Å. Eight subunits (four dimers) were identified in the asymmetric unit. In addition to the two zinc ions located at the enzyme active site, a number of peaks of strong electron density were modelled as zinc or sulfate ions. The polypeptide chain could be modelled from residue 26 to residue 414, adding or remove one or two residues from either termini on some chains. In addition, residues 330-331 in chain D, 256-257 in chain F, and 355 in chain G were not visible in the electron density. There was not significant perturbation in the overall structure of the protein as compared to the published structure of CPG2 (PDB ID 1CG2<sup>12</sup>), with RMSDs ranging from 0.7 – 1.3 Å when comparing the four subunits of the 1CG2 crystal structure to the eight subunits in the CPG2-BMIE- $\text{Zn}^{2+}$  structure. The active sites also appeared to be preserved in the new crystal structure, as compared to 1CG2.

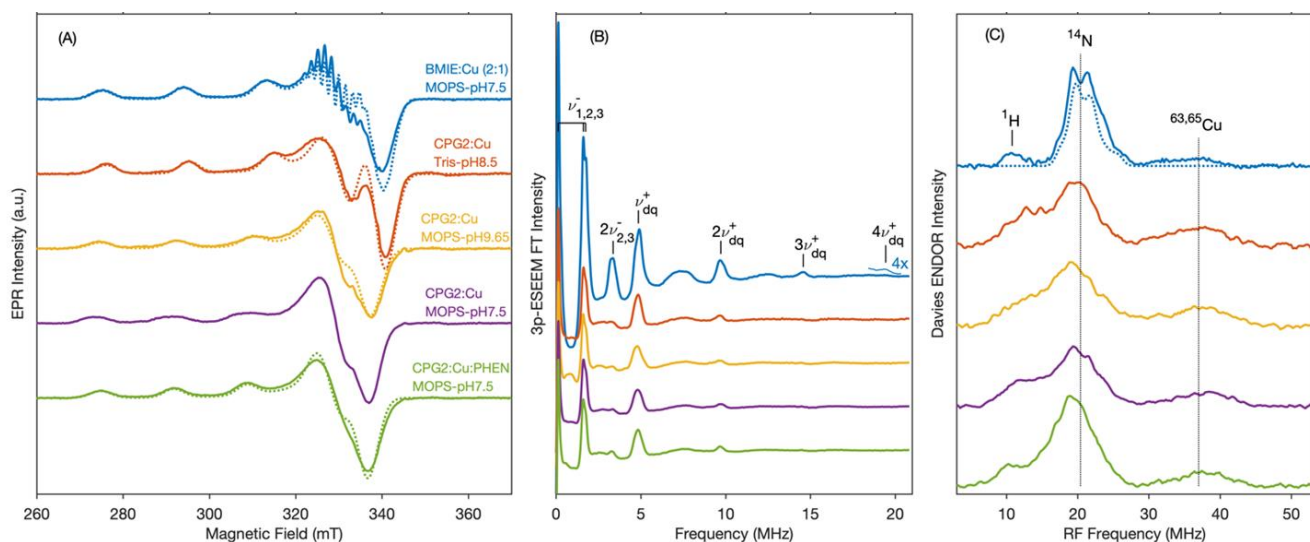


Figure 39-5. Figure 3. EPR characterization of binary and ternary BMIE:Cu complexes.

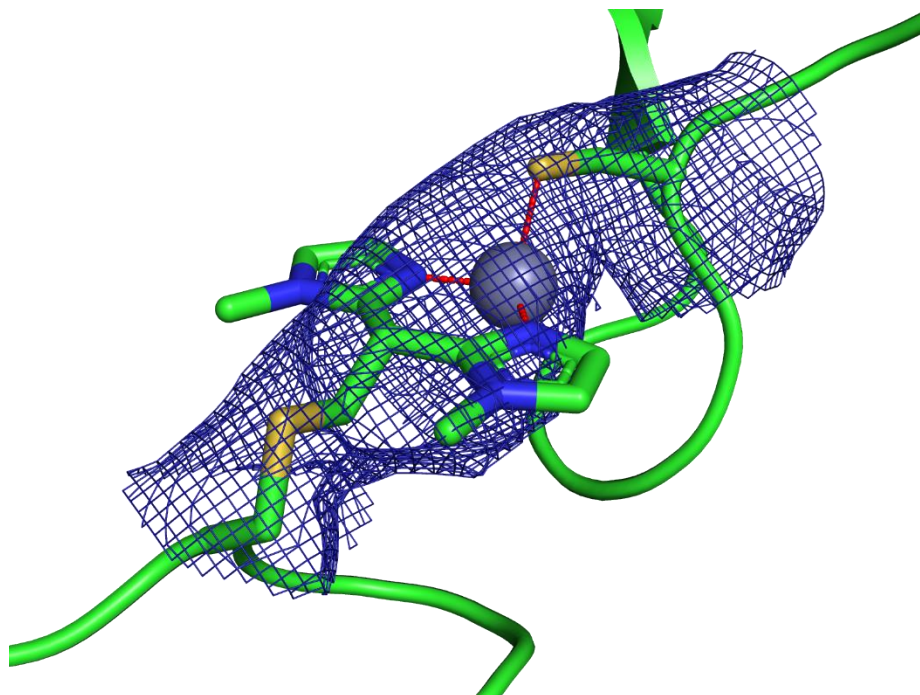
Each panel (A-C) shows the spectra for five samples: (top to bottom) Small-molecule BMIE:Cu (2:1) complex in MOPS buffer at pH 7.5; binary CPG2-BMIE:Cu cluster in Tris buffer at pH 8.5; binary CPG2-BMIE:Cu cluster in MOPS buffer at pH 9.65 and pH 7.5; ternary CPG2-BMIE:Cu:PHEN cluster in MOPS buffer at pH 7.5 (see labels on panel A). The  $\text{Cu}^{2+}$  complex concentrations in all samples were in the range 100-200 mM. (A) Continuous wave EPR spectra measured at 20 K (solid lines) and simulated EPR spectra using the parameters from Table X (dashed lines). The EPR spectrum of CPG2-MOPS:Cu at pH 7.5 consists of multiple Cu species and therefore it was not simulated. (B) Three-pulse ESEEM spectra at 22 K measured with  $t = 204$  ns at magnetic field 338 or 340 mT ( $g^{\wedge}$  orientation). The basic  $^{14}\text{N}$  nuclear spin transitions ( $\nu_i^-$  and  $\nu_{dq}^+$ ) and their harmonics ( $n \cdot \nu_{dq}^+$ ) for amine  $\text{CH}_3$ -nitrogen of coordinated BMIE are labeled. (C) Davies ENDOR spectra at 22 K measured using non-selective microwave pulses ( $p = 32$  ns) at magnetic field 338-340 mT ( $g^{\wedge}$  orientation). Nuclear spin transitions from central  $^{63,65}\text{Cu}$ , directly-coordinated imine  $^{14}\text{N}$  of BMIE and strongly-coupled  $^1\text{H}$  protons are labeled. The dashed line shows the simulated ENDOR spectrum for BMIE:Cu (2:1) using  $^{14}\text{N}$  hyperfine parameters from Table 1.

Table 9-5. Table 1. EPR hyperfine couplings and g-factors

Sample	Buffer	Electron g-factor ( $g_x, g_y, g_z$ )	Hyperfine couplings ( $A_x, A_y, A_z$ ) MHz	
			$^{63}\text{Cu}$	Imine $^{14}\text{N}$ of BMIE
BMIE:Cu (2:1)	MOPS/pH 7.5	(2.046, 2.052, 2.2365)	(72, 57, 580)	(40, 46, 39)
CPG2-BMIE:Cu	Tris/pH 8.5	(2.055, 2.045, 2.226)	(43, 57, 589)	
CPG2-BMIE:Cu	MOPS/pH 9.65	(2.064, 2.054, 2.253)	(43, 57, 546)	
CPG2-BMIE:Cu:PHEN	MOPS/pH 7.5	(2.064, 2.054, 2.261)	(43, 57, 525)	

Residue Ser203 had been substituted to Cys in order to accommodate covalent attachment of a BMIE molecule to the free Cys thiol group. The eight Cys203 residues in the asymmetric unit all occur at crystallographic subunit interfaces, with Cys203 from one subunit in close proximity to the Cys203 of the neighboring subunit ( $C\alpha$ - $C\alpha$  distance of 12 Å, similar to those residues in the crystallization interface of 1CG2). A large peak in the electron density map between the two Cys residues was modelled as a BMIE-modified Cys residue coordination  $Zn^{2+}$ . Interestingly, the density appeared to provide space for only a single BMIE-modified Cys203 per interface, with the neighboring subunit's Cys203 appearing to form an interaction with the BMIE-coordinated zinc ion. The lack of strong density for the entirety of the methyl-imidazole moieties may suggest a certain amount of rotational flexibility.

Examination of the electron density map at and around the BMIE-Cys203 sites show diffuse density, which, along with the quality of density, suggests that there may be some flexibility in this region. We cannot exclude that in some asymmetric units, the modified and unmodified cysteines are switched. It is also possible that in a minority of cases, both Cys203 are modified with one of the BMIE-Cys203s rotated away from the interface. At this resolution, however, the current model provides the best agreement to the electron density.



**Figure 40-5.** Figure 4. Electron density at the Cys203-BMIE-Zn<sup>2+</sup> site. The 2F<sub>o</sub>-F<sub>c</sub> 3.0 Å from the sidechain atoms of residue 203 (both subunits) is shown contoured at 0.9 σ. The coordination of atoms with the Zn<sup>2+</sup> ion is shown as red dashed lines.

**Table 10-5 Table 2:** Crystallographic data for [Zn(BMIE-CPG2-S203C)]<sup>2+</sup>

Data collection statistics	
Space group	P2 <sub>1</sub>
Number of Molecules per Asymmetric Unit	8
a (Å)	72.12
b (Å)	188.46
c (Å)	124.50
α = γ (°)	90.00
β (°)	90.17
Wavelength (Å)	1.0322
Resolution Range (Å) <sup>†</sup>	41.50 – 3.11 (3.16 – 3.11)
Completeness (%) <sup>†</sup>	99.87 (99.93)
Redundancy <sup>†</sup>	6.64 (6.91)
I/σ	4.3 (1.1)
R <sub>merge</sub> (%) <sup>†</sup>	0.389 (1.949)
R <sub>meas</sub> (%) <sup>†</sup>	0.423 (2.106)
R <sub>pim</sub> (%) <sup>†</sup>	0.164 (0.793)
CC <sub>1/2</sub> <sup>†</sup>	0.977 (0.358)
Refinement statistics	
Total number of reflections (reflections in R-free	56,534 (2981)

R <sub>factor</sub> (%)	0.274
R <sub>free</sub> (5% free test set) (%)	0.309
Number of atoms	21,542
Protein	21,359
Water	121
Catalytic Zn <sup>2+</sup>	16
Buffer components	46
RMSD	
Bond length (Å)	0.003
Bond angle (°)	1.142
Average atomic B-Factor (Å <sup>2</sup> )	72.37
Protein (Å <sup>2</sup> )	72.59
Water (Å <sup>2</sup> )	32.11
Buffer components (Å <sup>2</sup> )	60.88
Inhibitor (Å <sup>2</sup> )	82.82
Ramachandran Plot	3081
Residues in Favoured Positions	2816
Residues in Allowed Positions	247
Residues in Disallowed Positions	18

---

<sup>†</sup>Items in parentheses refer to the highest resolution shell.

## 5.5 Conclusion

For this work, we sought to develop a compact site-selective protein modification conjugate that would enable diverse metal binding. Although bis(1-methyl-1H-imidazole-2-yl)ethene or BMIE has been previously reported, its characterization as a bioconjugate had yet been explored. We improved the synthesis of BMIE with Martin's Sulfurane. Small molecule adduct characterization confirmed thiol conjugation and metal binding, a diverse divalent metal-binding palette (Co, Ni, Cu, and Zn), and versatile chelation (N<sub>2</sub> and N<sub>2</sub>S\*). Modification of S203C substituted Carboxypeptidase G2 (>90%) confirmed BMIE utility as a biocompatible conjugate addition. Cu and Zn metal coordination by the modified protein was qualitatively measure with ESI-MS and ICP-MS. Coordination stoichiometry (1:1) and geometry (tetragonal) of the Cu:BMIE-CPG2-S203C species

were measured using EPR. We observed unique Cu:BMIE-CPG2-S203C species in the presence of various counter ligands (H<sub>2</sub>O, Tris, and Phenanthroline), which suggests BMIE-modified proteins will offer versatility at labile coordination positions for future Cu-centered chemistry. A crystal structure of modified CPG2 co-crystallized with ZnCl<sub>2</sub> revealed a novel Zn:BMIE coordination interaction across the crystallographic interface of each monomer in the asymmetric unit. This result further supports the potential for Zn-coordination protein design with BMIE. These features, combined with ease of synthesis, make for an attractive metalloprotein design tool and should enable future catalytic and structural applications.

## 5.6 Supplemental Information

### 5.6.1 Experimental Procedures

#### 5.6.1.1 Synthesis procedures

**1,1-Bis(1-methyl-1H-imidazol-2-yl)ethan-1-ol.** A solution of 1-methylimidazole (0.8 ml, 10 mmol) in 25 ml of freshly distilled THF was stirred under a nitrogen atmosphere at -78 °C (Dry Ice bath temperature). A solution of *n*-BuLi (5.7 mL, 9.14 mmol, 1.6 M in hexane) was added dropwise over 10 min. After 5 min, the cooling bath was removed and the reaction mixture was allowed to warm to room temperature. After 30 min, the reaction mixture was again cooled to -78 °C. Ethyl acetate (0.3 mL, 3 mmol) was added in one portion. After 5 min, the cooling bath was removed, and the reaction mixture was allowed to warm to room temperature. After 3 h, the reaction mixture was quenched with saturated aq ammonium chloride, and then concentrated. The residue was partitioned between aq saturated sodium bicarbonate and dichloromethane. The aqueous layer was washed twice more with dichloromethane, and the combined organic extract was dried over anhydrous sodium sulfate and then concentrated. The residue was dissolved in a minimum amount of chloroform. Hexane was added by drops to the cloud point, and the mixture was stored at -20 °C. The white crystalline product was collected by filtration (0.378 gm, 60%): mp 165

– 169 °C;  $^1\text{H}$  NMR (500 MHz,  $\text{CDCl}_3$ )  $\delta$  6.89 (d,  $J$  = 1.15 Hz, 2H), 6.75 (d,  $J$  = 1.15 Hz, 2H), 6.21 (br s, 1H), 3.23 (s, 6H), 2.01 (s, 3H);  $^{13}\text{C}$  NMR (125 MHz,  $\text{CDCl}_3$ )  $\delta$  28.01, 33.71, 69.60, 123.58, 125.80, 148.66; ESI-MS  $[\text{M}+\text{H}]^+$   $m/z$  calcd for  $\text{C}_{10}\text{H}_{15}\text{N}_4\text{O}$ ; 207.12 found 207.1

**2,2'-(Ethene-1,1-diyl)bis(1-methyl-1H-imidazole), “BMIE.”** A solution of Martin’s sulfurane (0.45 g, 0.67 mmol) in 2 mL of dichloromethane was added slowly under an atmosphere of nitrogen to a stirred solution of 1,1-bis(1-methyl-1H-imidazol-2-yl)ethanol (0.069 g, 0.335 mmol) in 2 mL of dichloromethane. After 24 h, the solution was concentrated to a pale-yellow oil. The crude product was purified by chromatography on silica, eluting with 98:1 dichloromethane / methanol, and then 8:1:0.3:0.15 ethyl acetate / methanol / ammonium hydroxide / water, to give the title compound (40 mg, 62%) as a yellow oil:  $^1\text{H}$  NMR (500 MHz,  $\text{CDCl}_3$ )  $\delta$  7.07 (d,  $J$  = 1.2 Hz, 2H) and 6.88 (d,  $J$  = 1.15 Hz, 2H), 5.99 (s, 2H), 3.29 (s, 6H);  $^{13}\text{C}$  NMR (125 MHz,  $\text{CDCl}_3$ ):  $\delta$  33.40, 122.42, 124.30, 128.67, 136.64, 154.74; ESI-MS  $[\text{M}+\text{H}]^+$   $m/z$  calcd for  $\text{C}_{10}\text{H}_{13}\text{N}_4$ ; 189.11 found 189.1

**Methyl S-(2,2-Bis(1-methyl-1H-imidazol-2-yl)ethyl)-N-(tert-butoxycarbonyl)-L-**

**cysteinate.** A solution of N-(tert-butoxycarbonyl)-L-cysteine methyl ester (200 mg, 1.275 mmol) in 1 mL of methanol was stirred in a vial at room temperature. A solution of BMIE (60 mg, 0.319 mmol) in 500  $\mu\text{L}$  of methanol was added in one portion. The reaction was allowed to stir for 2 h, and then was concentrated to a residue.

Chromatography on silica with 70:5:1.5 ethyl acetate / methanol / ammonium hydroxide as the eluant gave the BMIE-cysteine adduct as a yellow oil that solidified upon standing (65 mg, 48%):  $^1\text{H}$  NMR (300 MHz,  $\text{CD}_3\text{CN}$ )  $\delta$  = 6.87–6.90 (m, 2H), 6.85 (br app t, 1H,  $J$  = 0.6 Hz), 6.83 (br app t,  $J$  = 0.6 Hz, 1H), 6.42 (br d,  $J$  = 5.1 Hz, 1H), 4.54 (t,  $J$  = 4.6 Hz, 1H), 4.35 (br q,  $J$  = 4.8 Hz, 1H), 3.66 (s, 3H), 3.45 (s, 3H), 3.43 (s, 3H), 3.37–3.43 (partially obscured m, 2H), 2.84–2.86 (m, 2H), 1.41 (s, 9 H);  $^{13}\text{C}$  NMR (125 MHz,

CD<sub>3</sub>CN)  $\delta$  171.6, 155.6, 145.72, 145.68, 126.6, 126.5, 122.0, 121.9, 79.1, 53.9, 51.9, 37.9, 34.4, 33.8, 32.22, 32.18, 27.6; LR-MS-ESI [M+1]<sup>+</sup> calcd for C<sub>19</sub>H<sub>30</sub>N<sub>5</sub>O<sub>4</sub>S<sup>+</sup> 424.2; found 424.1.

**Crystallizations of metal complexes: with copper(II) chloride.** The BMIE-thiocresol adduct, 2,2'-(2-(p-tolylthio)ethane-1,1-diyl)bis(1-methyl-1H-imidazole) (5 mg, 1 equiv), was dissolved in 1 mL of acetonitrile in a small vial. Copper(II) chloride (2.13 mg, 1 equiv) was added. The vial was covered loosely and acetonitrile was allowed to evaporate over a few days to afford blue-green crystals. Crystallographic analysis indicated a 1:1 complex.

**With copper(II) trifluoromethanesulfonate.** The BMIE-thiocresol adduct, 2,2'-(2-(p-tolylthio)ethane-1,1-diyl)bis(1-methyl-1H-imidazole) (5 mg, 1 equiv), was dissolved in 1 mL of acetonitrile in a small vial. Copper(II) triflate (5.71 mg, 1 equiv) was added. The vial was covered loosely and acetonitrile was allowed to evaporate over a few days to afford light violet crystals. Crystallographic analysis indicated a 2:1 adduct/Cu(II) complex.

**With zinc(II) chloride.** The BMIE-thiocresol adduct, 2,2'-(2-(p-tolylthio)ethane-1,1-diyl)bis(1-methyl-1H-imidazole) (2 mg, 1 equiv), was dissolved in 1 mL of degassed acetonitrile in a small vial. The vial was placed inside a larger vial, purged with argon, and then ether was added to the larger outside vial. Zinc(II) chloride (0.87 mg, 1 equiv, 10 mg/mL) solution in acetonitrile was added to the smaller vial. The larger vial was capped tightly, and ether was allowed to diffuse overnight to afford crystals. Crystallographic analysis indicated a 1:1 complex.

**With tetrakis(acetonitrile) copper(I) tetrafluoroborate.** The BMIE-thiocresol adduct, 2,2'-(2-(p-tolylthio)ethane-1,1-diyl)bis(1-methyl-1H-imidazole) (2 mg, 1 equiv), was dissolved in 1 mL of degassed acetonitrile in a small vial. The vial was placed inside a larger vial and purged with argon. Degassed diethyl ether was added to the larger vial. Tetrakis(acetonitrile) copper(I) tetrafluoroborate (200 mL of a 10 µg/mL solution in acetonitrile, 1 equiv) was added to the smaller vial. The larger vial was capped tightly, and ether was allowed to diffuse overnight to afford crystals.

**NMR-monitored titrations: with zinc(II) trifluoromethanesulfonate.** The BMIE-cysteine adduct, methyl S-(2,2-bis(1-methyl-1H-imidazol-2-yl)ethyl)-N-(tert-butoxycarbonyl)-L-cysteinate (10 mg, 0.023 mmol, 1 equiv), was dissolved in 0.4 mL of acetonitrile-d<sub>3</sub> in an NMR tube. Zinc(II) triflate (21.5 µL of a 200 mg/mL solution in acetonitrile-d<sub>3</sub>, 4.29 mg, 0.5 equiv) was added, and after 5 min the <sup>1</sup>H NMR spectrum was recorded. Serial addition of 0.5 equiv of zinc(II) triflate was repeated up to 2.5 total equiv, with the <sup>1</sup>H NMR spectrum recorded after each addition.

**With zinc(II) chloride.** A solution of the BMIE-cysteine adduct, methyl S-(2,2-bis(1-methyl-1H-imidazol-2-yl)ethyl)-N-(tert-butoxycarbonyl)-L-cysteinate (5 mg, 1 equiv), in acetonitrile-d<sub>3</sub> (0.4 mL) was added to an NMR tube. A solution of zinc(II) chloride (0.8 mg, 0.5 equiv) in 35 µL of acetonitrile-d<sub>3</sub> was added, and after 5 min the <sup>1</sup>H NMR spectrum was recorded. Serial addition of 0.5 equiv of zinc(II) chloride was repeated up to 2.5 total equiv, with the <sup>1</sup>H NMR spectrum recorded after each addition.

**With tetrakis(acetonitrile) copper(I) tetrafluoroborate.** Ascorbic acid (10 mg) was added to an NMR tube, and the tube was purged three times with argon. A solution of the BMIE-cysteine adduct, methyl S-(2,2-bis(1-methyl-1H-imidazol-2-yl)ethyl)-N-(tert-

butoxycarbonyl)-L-cysteinate (5 mg, 1 equiv), in 0.4 mL of acetonitrile- $d_3$  was added to the NMR tube. A solution of tetrakis(acetonitrile) copper(I) tetrafluoroborate (1.86 mg, 0.5 equiv) in 35  $\mu$ L of acetonitrile- $d_3$  was added and after 5 min the  $^1\text{H}$  NMR spectrum was recorded. The argon atmosphere was maintained. Serial addition of 0.5 equiv of tetrakis(acetonitrile) copper(I) tetrafluoroborate was repeated up to 2.5 total equiv, with the  $^1\text{H}$  NMR spectrum recorded after each addition.

#### 5.6.1.2 Protein Expression and Purification

Cells were grown at 37°C (shaking) in auto-induction media containing: 5g of tryptone, 2.5g of yeast extract, 0.1% glycerol, 0.01% glucose, 0.1%  $\alpha$ -lactose, 0.1M  $(\text{NH}_4)_2\text{SO}_4$ , 0.2M  $\text{KH}_2\text{PO}_4$ , 0.2M  $\text{Na}_2\text{HPO}_4$ , 1mM  $\text{MgSO}_4$ , 1mM trace metal mix. After 3 hours the temperature was dropped to 18°C where it remained for 24 hours.

Cell cultures were pelleted at 4K rcf for 45 minutes. Cell pellets were resuspended in 30ml of 50mM Tris, 250mM NaCl, 25mM imidazole, pH 7.4 and sonicated. Sonicated lysate was centrifuged again at 40K rcf for 45 min to remove cell debris. Lysate was collected and run over a 0.22 $\mu$ M filter before running over a standard nickel affinity column (NTA beads and 100mM  $\text{NiSO}_4$ ). Column was washed with excess resuspension buffer and eluted with 50mM Tris, 250mM NaCl, 250mM imidazole, pH 7.4. 20ml elution volumes typically produced 5-10mg/ml of CPG2.

TEV protease (with his tag) was added to purified CPG2 solution (1/100) and was dialyzed into 50mM Tris-HCl, 0.5mM EDTA, 1mM DTT, pH 8.0 at 4C overnight.

Followed by dialysis back into resuspension buffer (see purification) before being place

back on Ni-NTA column. Flow through was collected along with 10ml of wash (resuspension).

#### 5.6.1.3 Protein Characterization.

*Protein labeling:* Proteins were exchanged into Chelexed Ellman's reagent buffer 100mM Na<sub>2</sub>PO<sub>4</sub>, 1mM EDTA, pH 8.0 using PD10 desalting columns. BMIE was added to a final concentration of 2mM. Protein was left to react for 4 hours at room temperature followed by 18-24 hours at 4C.

*Ellman's assay:* Proteins were exchanged into Ellman's reagent buffer (see protein labeling), to each sample or at each time point, DTNB was added to a final concentration of 20uM, sample was left to react at room temperature for 30 minutes before measuring absorbance at 412nm.

*ICP-MS.* Proteins were exchanged into metal-chelation buffer (10mM Tris, 100mM NaCl, pH 8.0 passed over filter after 30 minute incubation with Chelex-100 at room temperature). 2ml of sample was measured for each condition. Protein samples had a final concentration of 20-25uM and 3-5 eqv. of metal salt (depending on the sample). A linear standard was made by measuring out l-cysteine, ZnCl<sub>2</sub>, and CuCl<sub>2</sub> and creating a ½ serial dilution over 5 samples to relate metal count to concentration.

*Electron Paramagnetic Resonance (EPR) spectroscopy:* All EPR experiments were carried out using a Bruker EPR spectrometer (E580e) operating at X-band microwave frequency. Helium-flow cryostats (Oxford ESR900 and CF935) equipped with an Oxford temperature controller (ITC503) were used for cryogenic temperatures.

*Continuous wave (CW) EPR experiments* of  $\text{Cu}^{2+}$  complexes were performed at temperature 20 K with the following experimental settings: microwave frequency, 9.49 GHz; microwave power, 200  $\mu\text{W}$ ; modulation amplitude, 0.5-1 mT. Concentration of  $\text{Cu}^{2+}$  clusters in each sample was determined by comparing the measured  $\text{Cu}^{2+}$  signal intensity to the EPR standard with a known number of spins (a  $\text{CuSO}_4 \cdot 5\text{H}_2\text{O}$  crystal of known weight in mineral oil).

CW EPR simulations were performed using the EasySpin toolbox for MATLAB (<http://www.easyspin.org/>).<sup>1</sup> The electron spin g-factor tensors and copper hyperfine coupling tensors resulting from the spectral simulations are summarized in **Table X**.

*Pulsed EPR experiments*, including two-pulse ESEEM ( $\pi/2 - \tau - \pi - \tau - \text{echo}$ ), three-pulse ESEEM ( $\pi/2 - \tau - \pi/2 - T - \pi/2 - \tau - \text{echo}$ ) and Davies ENDOR,<sup>2-4</sup> were performed to characterize nuclear spin environment of coordinated  $\text{Cu}^{2+}$  complexes. An appropriate phase cycling was used in each pulsed experiment to eliminate contributions from unwanted echoes.<sup>2</sup>

Prior to Fourier transformation (FT), the ESEEM time-domains were baseline corrected by fitting the oscillating ESEEM decay with a stretched exponential decay function, dividing the experimental decay by the fit function and subtracting a unity. This baseline correction procedure resulted in FT ESEEM spectral intensities which were automatically normalized(!) to a unit echo signal amplitude. This normalization procedure allowed us to directly compare spectral amplitudes in ESEEM spectra measured for different samples, with different  $\text{Cu}^{2+}$  concentrations, etc.<sup>3</sup>

After Fourier transformation, linear phase correction was applied to the FT spectra in order to correct for missing dead times. In case of two-pulse ESEEM, the dead time ( $t_0$ ) was determined by initial  $\tau$  delay between the  $\pi/2$  and  $\pi$  pulses in the experiment. In case of three-pulse ESEEM, the corrected time was calculated as  $(\tau + t_0)$ , where  $\tau$  is the fixed delay between the first  $\pi/2$  and second  $\pi/2$  pulses in the three-pulse sequence, and  $t_0$  is the initial delay between the second and third pulses. All (two-pulse and three-pulse) ESEEM spectra reported in this work are normalized, phase-corrected cosine Fourier-Transforms.

Davies ENDOR (Electron Nuclear Double Resonance) spectra were measured with the sequence  $(\pi - T - \pi/2 - \tau - \pi - \tau - \text{echo})$ , where a radio frequency  $\pi$  pulse (6  $\mu\text{s}$ ) was applied during T period. Non-selective microwave pulses ( $\pi/2$  pulse = 16 ns, and  $\pi$  pulse = 32 ns) were used to suppress a contribution from weakly-coupled  $^1\text{H}$  protons. To allow a comparison of ENDOR signal intensities measured on different samples the ENDOR amplitude was normalized to the echo intensity off-resonance with RF transitions.

Other pulsed EPR settings: microwave frequency, 9.687 GHz; magnetic field, 285-287 mT (the  $g_{\parallel}$  field orientation) and 338-342 mT (the  $g_{\perp}$  field orientation); microwave  $\pi/2$  and  $\pi$  pulses, 16 and 32 ns, respectively; initial  $\tau$  delay, 140 ns; integration window, 16 ns (ESEEM) and 60 ns (Davies ENDOR); shot repetition times, 1-2 ms; and temperature, 22 K.

*Crystallization, data collection, and refinement of CPG2-BMIE crystal structure:*

Following gel filtration, CPG2-BMIE was diluted 5 mg/mL in 50 mM Tris 100 mM NaCl pH 7.4 supplemented with 100 mM  $\text{ZnSO}_4$  for crystallization. Crystals were obtained

using hanging drop vapor diffusion. A 2  $\mu$ L drop of 5 mg/mL protein sample was mixed with 2  $\mu$ L of reservoir solution and suspended over the reservoir containing 0.2 M Tris pH 7.5, 10% PEG 3350, and 5% glycerol. The crystallization trays were incubated at 4°C, and crystals were obtained in 4-20 days.

Crystals were harvested and flash-frozen in liquid nitrogen. Data were collected under standard cryogenic conditions at the GM/CA@APS beamline 23ID-D. Data were processed using xia2<sup>5</sup> with the XDS and XSCALE<sup>6</sup> pipeline, indexing with peaks from all images, in space group P2<sub>1</sub>. The structure was solved by molecular replacement using Phaser,<sup>7</sup> with the 1CG2 crystal structure<sup>8</sup> being used as a search model. Eight CPG2 monomers were found per asymmetric unit.

The crystal structure was refined using multiple rounds of positional and B-factor refinement using Refmac<sup>9</sup> with non-crystallographic restraints. Coot<sup>10</sup> was used to regularly perform manual model building during refinement. The final coordinates and structure factors have been deposited to the Protein Data Bank as PDB ID XXXX.

## 5.6.2 Supplemental Information

### 5.6.2.1 Small-molecule BMIE:Cu (2:1) complex

The CW EPR spectrum of BMIE:Cu (2:1) (see **Fig. 3(A)** in the main text) is characteristic of an elongated octahedral (or square planar) geometry of a Cu<sup>2+</sup> complex with two BMIE ligands symmetrically coordinated in the equatorial plane. The copper hyperfine peaks resolved at  $g_{\parallel}$  are symmetric in shape (no shoulders), thus confirming the unique speciation of the BMIE:Cu (2:1) cluster. Nine <sup>14</sup>N hyperfine-split lines resolved at

$g_{\perp}$  orientation (330 mT) further support the unique cluster speciation with four equivalent  $^{14}\text{N}$  nuclei (two symmetric BMIE ligands) coordinated in an equatorial plane. The  $g$ -tensor parameters and hyperfine coupling parameters for copper and four nitrogen nuclei extracted from the spectral simulation (the dashed line in **Fig. 3(A)**) are collected in **Table 1**.

The echo-detected EPR of BMIE:Cu (2:1) (**Fig. S10**, top) shows the spectral lineshape that is similar to the CW EPR, although now presented in the absorption mode as compared to the first-derivative shape of the CW EPR spectrum. Two principal field positions (marked with solid circles in **Fig. S10**) were used for all ESEEM and ENDOR experiments reported in this work. At these field positions only small sub-populations of Cu clusters, whose principal symmetry axis ( $g_z$ ) is parallel or perpendicular to the applied magnetic field, are selectively excited and measured in the ESEEM/ENDOR. These principal field positions are known as the  $g_{\parallel}$  and  $g_{\perp}$  field orientation positions.<sup>2,3</sup>

Two-pulse ESEEM decay for BMIE:Cu (2:1) measured at the  $g_{\perp}$  field orientation (**Fig. S11(A)**). The deep-amplitude ESEEM oscillations are observed, characteristic of  $^{14}\text{N}$  nuclear spins with their hyperfine couplings close to the so-called "cancellation condition", e.g. when  $|A/2 - \nu_I| \ll K$ , where  $A$  is the  $^{14}\text{N}$  hyperfine coupling,  $\nu_I$  is the  $^{14}\text{N}$  Zeeman frequency, and  $K = e^2 Qq/4h$  is the  $^{14}\text{N}$  nuclear quadrupolar constant.<sup>11,12</sup> These oscillations show up as the low-frequency peaks (below 5 MHz) in the Fourier-transform (FT) spectrum in **Fig. S11(C)**. The spectrum is complex, showing many peaks with positive and negative amplitudes. To aid in interpretation of these peaks we performed three-pulse ESEEM experiments (see below). Thus, the peak assignments to

the basic transitions ( $\nu_{1,2,3}^-$  and  $\nu_{dq}^+$ ) and also various combination transitions, as shown in **Fig. S11(C)**, are based on the three-pulse ESEEM spectra.

Three-pulse ESEEM spectra (see **Fig. 3(B)** in the main text and also **Fig. S13** and **S14**) have several key advantages as compared to the two-pulse ESEEM spectra. First, the three-pulse ESEEM spectra are less crowded showing a fewer number of peaks, since only basic transitions (and their harmonics) are observed and no combination peaks present. Second, the three-pulse ESEEM spectra demonstrate narrower peak linewidths and therefore the enhanced spectral resolution (a simple consequence of a longitudinal relaxation time  $T_1$  being much longer than a transverse time  $T_2$ ). The enhanced spectral resolution is evident in resolving the single peak at  $\nu_{2,3}^- = 1.69$  MHz in the two-pulse ESEEM spectra into two distinct peaks at  $\nu_2^- = 1.62$  and  $\nu_3^- = 1.75$  MHz in the three-pulse ESEEM spectra. It is further evident in the enhanced peak intensity of the line at  $\nu_1^- = 0.13$  MHz. Lastly, the delay  $\tau$  in our three-pulse ESEEM experiment was selected such that to completely suppress the overlapping peak from remote  $^1\text{H}$  protons and to allow an unobscured detection of the third harmonic  $^{14}\text{N}$  peak at  $3 \cdot \nu_{dq}^+ = 14.6$  MHz.

Three peaks ( $\nu_{1,2,3}^-$ ) in the three-pulse ESEEM spectra follow a simple additive rule, e.g.  $\nu_3^- = \nu_1^- + \nu_2^-$ . These peaks can be assigned to the three basic  $^{14}\text{N}$  spin transitions for the electron spin projection "down". The peak at  $\nu_{dq}^+$  is then assigned to one of the three basic  $^{14}\text{N}$  spin transitions for the electron spin projection "up" (two other transitions at this electron spin projection are broad and unobservable). The remaining peaks in the ESEEM spectra are simple harmonics of  $\nu_{2,3}^-$  and  $\nu_{dq}^+$ .

Assuming closeness to the cancellation condition ( $|A/2 - \nu_I| \ll K$ ), the peak positions of the basic transitions is described by:

$$\nu_1^- = 2K\eta,$$

$$\nu_2^- = K(3 - \eta),$$

$$\nu_3^- = K(3 + \eta),$$

$$\nu_{dq}^+ = 2 \left[ \left( \frac{A}{2} + \nu_I \right)^2 + K^2(3 + \eta^2) \right]^{1/2},$$

where  $A$ ,  $\nu_I$  and  $K$  have been already defined above, and  $\eta$  is the asymmetry parameter of  $^{14}\text{N}$  nuclear quadrupolar tensor. Using these expressions, we can estimate a hyperfine coupling constant  $A = 2.36$  MHz and the  $^{14}\text{N}$  nuclear quadrupolar parameters  $K = 0.56 \pm 0.01$  MHz and  $\eta = 0.12 \pm 0.02$  (the  $^{14}\text{N}$  nuclear Zeeman frequency at 340 mT is  $\nu_I = 1.04$  MHz). Clearly, these derived parameters satisfy the "cancellation condition" ( $|A/2 - \nu_I| \ll K$ ), which validates our ESEEM pick assignments and the resulting  $^{14}\text{N}$  coupling estimates.

The estimated  $^{14}\text{N}$  nuclear quadrupolar parameters ( $K = 0.56 \pm 0.01$  MHz and  $\eta = 0.12 \pm 0.02$ ) are quite distinct and uniquely identify what type of  $^{14}\text{N}$  nitrogen is responsible for the observed ESEEM. Quadrupolar couplings of similar strength and asymmetry ( $K = 0.55 \pm 0.03$  MHz and  $\eta = 0.2 \pm 0.1$ ) have been reported for  $\text{Cu}^{2+}$  complexes with 1-methyl-imidazoles where a proton at the remote (amine) nitrogen of imidazole was substituted with a methyl group (see Table II in Jiang et al., 1990). The quadrupolar couplings for  $\text{N-CH}_3$  nitrogens were found distinctly different from amine

N-H nitrogens in unsubstituted imidazoles (or histidines)<sup>13</sup> and also different from any other type of <sup>14</sup>N nitrogen in non-imidazole, nitrogen-containing ligands.<sup>14</sup> Based on this comparison, we may conclude that amine N-CH<sub>3</sub> nitrogen(s) of BMIE ligands are responsible for our observed ESEEM spectra.

Four harmonics of the basic transition  $\nu_{dq}^+$  are clearly resolved in the three-pulse ESEEM spectra for BMIE:Cu (2:1) at the  $g_{\perp}$  field orientation (**Fig. 3(B)** in the main text and **Fig. S13**). These four harmonics indicate the presence of four(!) equivalent N-CH<sub>3</sub> nitrogens (two symmetrically-coordinated BMIE ligands) in BMIE:Cu (2:1).<sup>15</sup>

The three-pulse ESEEM spectra measured at the  $g_{\parallel}$  field orientation (**Fig. S13**, bottom) are similar to that at the  $g_{\perp}$  field orientation. From the peak positions in the  $g_{\parallel}$  spectra we estimate the hyperfine coupling constant  $A_{\parallel} = 2.07 \pm 0.02$  MHz and the <sup>14</sup>N nuclear quadrupolar parameters  $K = 0.57 \pm 0.02$  MHz and  $\eta = 0.16 \pm 0.03$ . The quadrupolar parameters are essentially the same, within the experimental error, at both  $g_{\parallel}$  and  $g_{\perp}$  field orientations as expected because of the closeness to the cancellation condition. However, the hyperfine coupling  $A_{\parallel} = 2.07$  MHz is noticeably smaller than  $A = 2.36$  MHz at the  $g_{\perp}$  orientations. The difference between two arises from an anisotropic hyperfine coupling contribution that is orientation-dependent and therefore different at the  $g_{\parallel}$  and  $g_{\perp}$  field orientations. From this difference we can estimate a magnitude of anisotropic hyperfine coupling  $T = 0.2$  MHz. In the dipole-dipole approximation, this anisotropic hyperfine coupling corresponds to a distance  $3.0 \pm 0.1$  Å between copper center and the amino N-CH<sub>3</sub> nitrogen of coordinated BMIE.

Strong hyperfine interactions with directly-coordinated (imine)  $^{14}\text{N}$  nitrogens of BMIE ligands can be probed using Davies ENDOR experiments (see **Fig. 3(C)** in the main text and **Fig. S15**). In these experiments we used non-selective microwave pulses ( $\pi$  pulse = 32 ns) to suppress overlapping signals from weakly-coupled  $^1\text{H}$  protons and to selectively measure only the transitions from strongly-coupled nuclei. The broad peak centered at 20 MHz is directly-coordinated (imine)  $^{14}\text{N}$  nitrogens of BMIE ligands showing a clear Zeeman splitting and a partially-resolved quadrupolar structure. The dashed line shows the simulated spectrum using the  $^{14}\text{N}$  hyperfine parameters for BMIE:Cu (2:1) from **Table 1** and assuming four symmetric  $^{14}\text{N}$  nuclei coordinated in the equatorial plane to  $\text{Cu}^{2+}$ . The quadrupolar couplings ( $K = 0.75$  MHz and  $\eta = 0.5$ ) were used in the simulation. The extracted  $^{14}\text{N}$  hyperfine parameters are similar to what reported previously for  $\text{Cu}^{2+}(\text{Imidazole})_4$  complexes.<sup>16</sup>

### 5.6.3 Supplemental References

1. Stoll, S. CW-EPR Spectral Simulations: Solid State. in *Methods in Enzymology* vol. 563 121–142 (Academic Press Inc., 2015).
2. Schweiger, A. (Arthur) & Jeschke, G. *Principles of pulse electron paramagnetic resonance*. (Oxford University Press, 2001).
3. Dikanov S. A., T. Y. D. *Electron Spin Echo Envelope Modulation (ESEEM) Spectroscopy*. (CRC Press, 1992).
4. Davies, E. R. A NEW PULSE ENDOR TECHNIQUE. *PHYSICS LETTERS* vol. 47.
5. Winter, G., Lobley, C. M. C. & Prince, S. M. Decision making in xia2. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **69**, 1260–1273 (2013).

6. Kabsch, W. Biological Crystallography. *Res. Pap. Acta Cryst* **66**, 125–132 (2010).
7. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
8. Rowsell, S. *et al.* *Crystal structure of carboxypeptidase G 2 , a bacterial enzyme with applications in cancer therapy.*
9. Murshudov Alexe, G. N., Vagin, A. & Dodson, E. J. *Refinement of Macromolecular Structures by the Maximum-Likelihood Method.* *Acta Cryst* vol. 53 (1997).
10. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 486–501 (2010).
11. Flanagan, H. L. & Singel, D. J. Analysis of <sup>14</sup>N ESEEM patterns of randomly oriented solids. *J. Chem. Phys.* **87**, 5606–5616 (1987).
12. Diiunov, S. A., Tsvetkov, Y. D. & Astashkc, A. V. *PARAMETERS OF QU~RUPOLE COUPLING OF <sup>14</sup>N NUCLEI IN CHLOROPHYLL *a* CATIONS DETERMINED BY THE ELETXRON SPIN ECHO METHOD.* vol. 90 (1982).
13. Jiang, F., Mccracken, J. & Peisach, J. *Nuclear Quadrupole Interactions in Copper(II)-Diethylenetriamine-Substituted Imidazole Complexes and in Copper(II) Proteins Table I. Ionization pKa and Apparent Binding Constant K (mM) of Model Compounds0 bases. J. Am. Chem. Soc* vol. 112 <https://pubs.acs.org/sharingguidelines> (1990).
14. Edmonds, D. T. *NUCLEAR QUADRUPOLE DOUBLE RESONANCE.*
15. McCracken, J. *et al.* *Electron Spin-Echo Studies of the Copper Binding Site in Phenylalanine Hydroxylase from Chromobacterium violaceum.* *Trav, Chim. Pays-Bays* vol. 110 <https://pubs.acs.org/sharingguidelines> (1988).

16. Iwaizumi, M., Kudo, T. & Kita, S. Bonding (Berlin) 1982, 51, 1. Downloaded via RUTGERS UNIV on March 23. 20, 11 (2020).

#### 5.6.4 Supplemental Figures

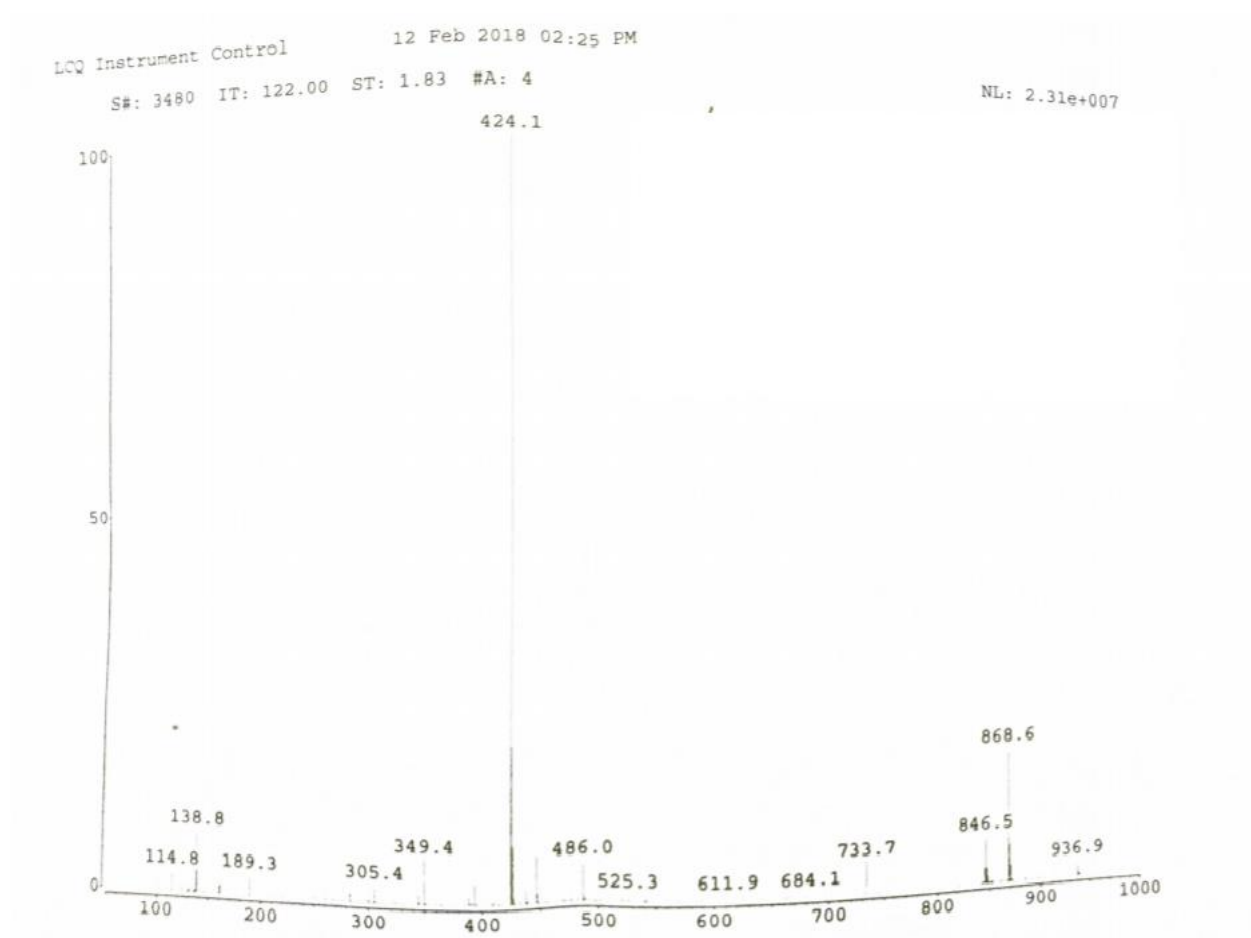


Figure 41-5. Figure S1. Mass spec date of BMIE-cys, expected MW 424 da.

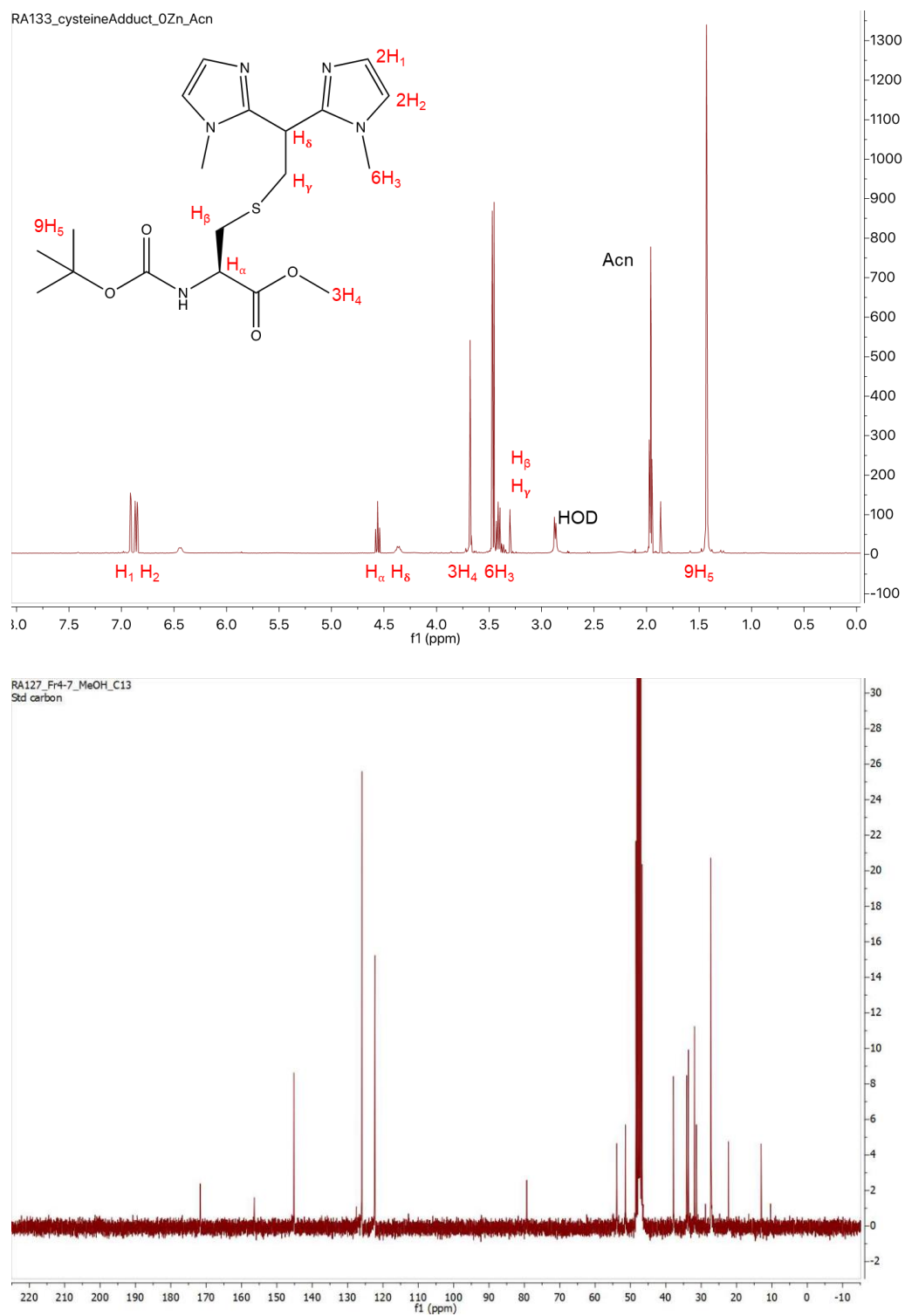


Figure 42-5. Figure S2.  $H^1$ -NMR and  $C^{13}$ -NMR of BMIE-cys

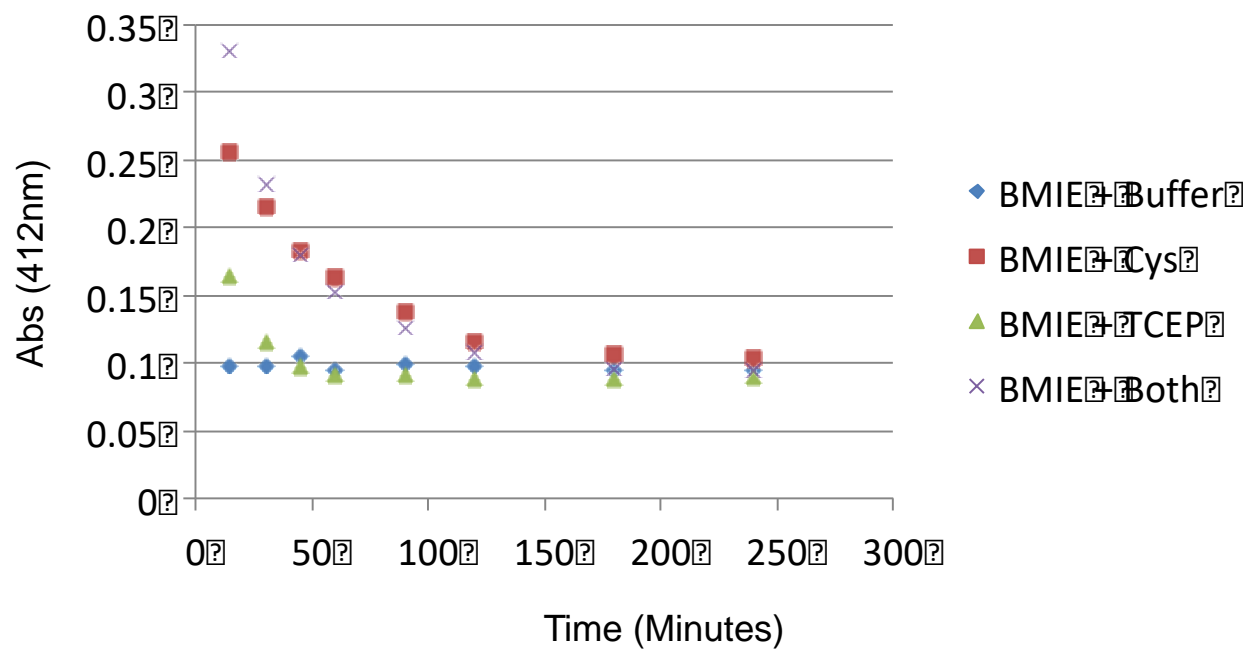


Figure 43-5. Figure S3. Ellman's Assay: BMIE-I-cys conjugation over time.

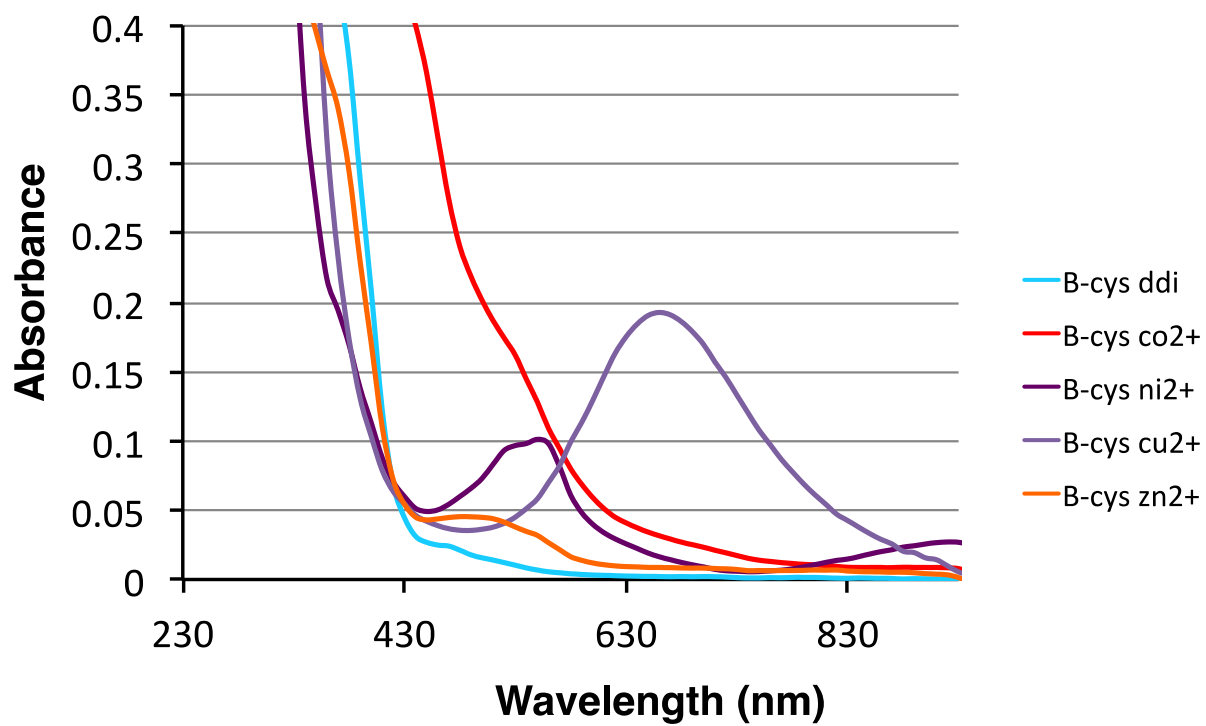
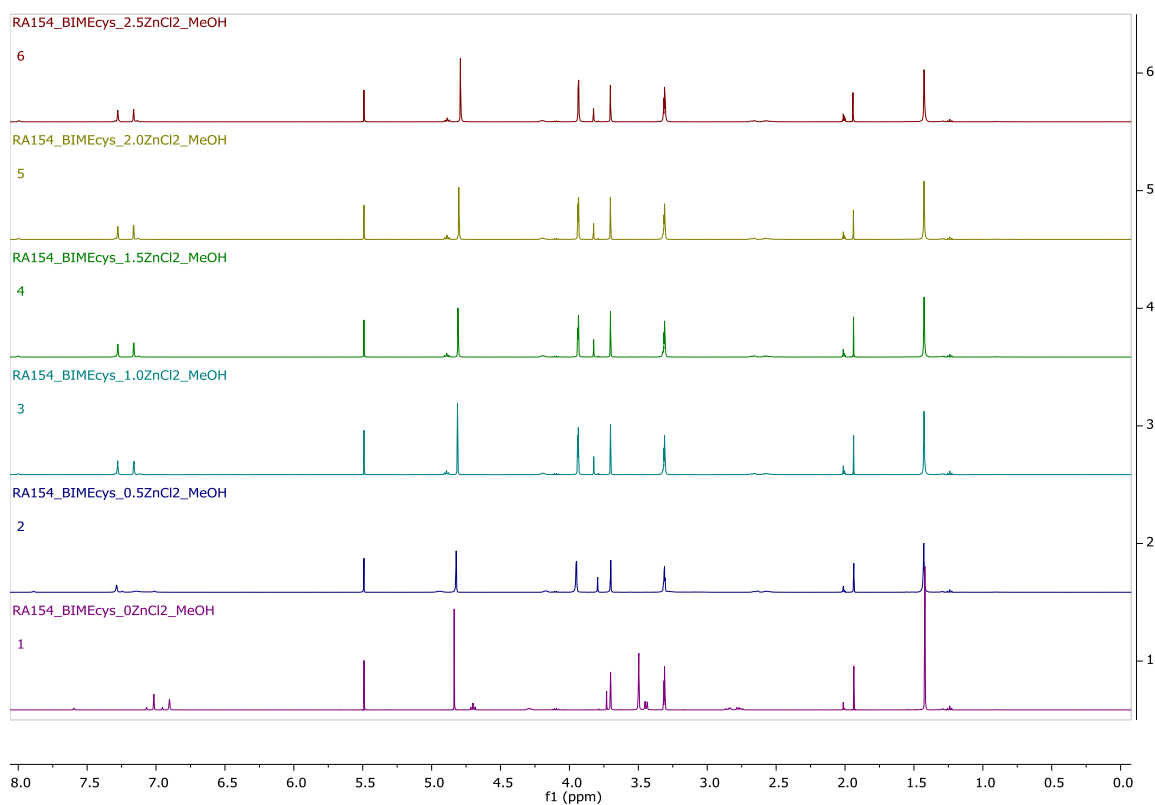
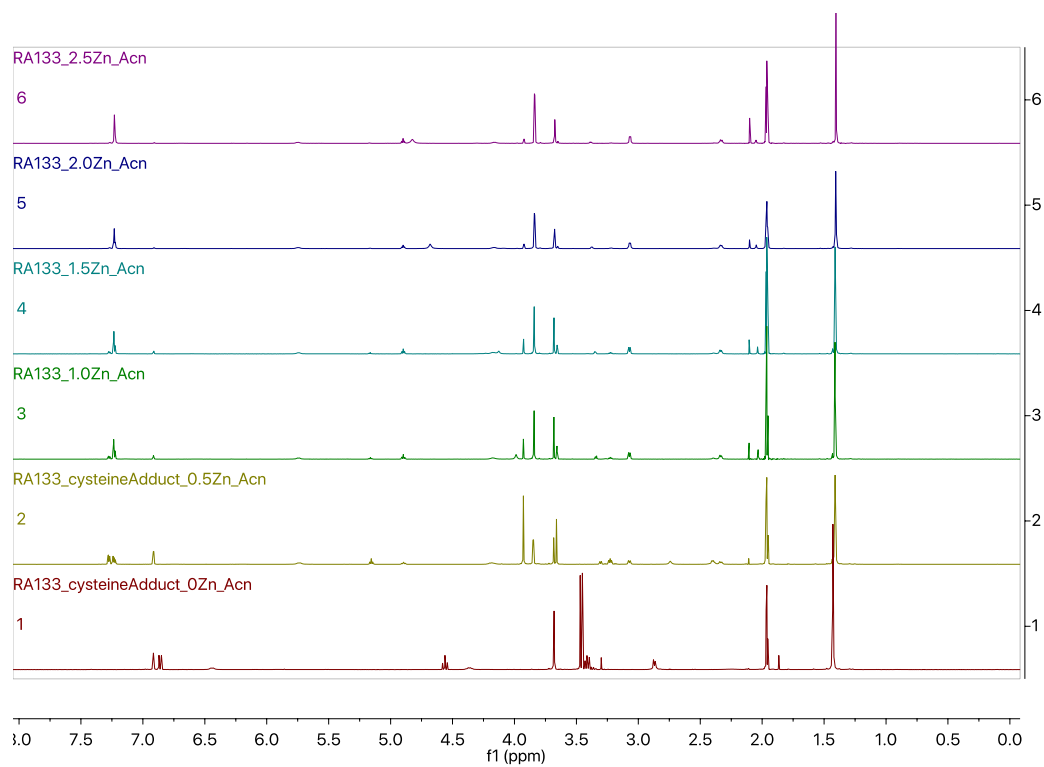


Figure 44-5. Figure S4. UV/Vis spectra of BMIE-l-cys and different divalent cations (Co, Ni, Cu, and Zn).



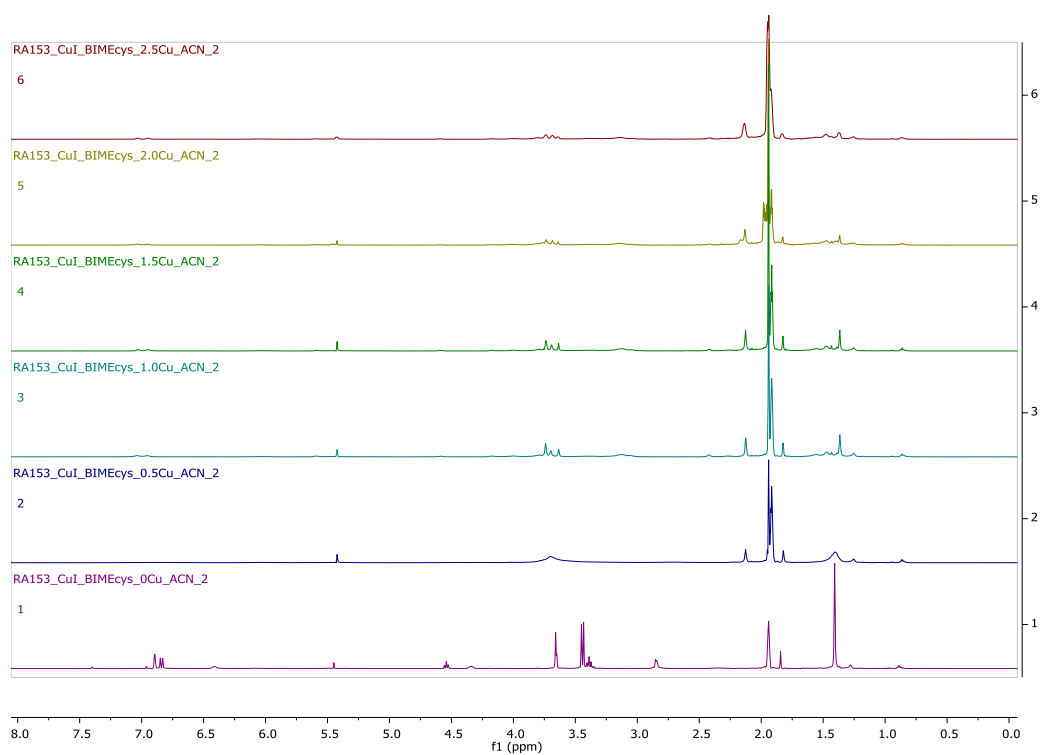
**Figure 45-5.** Figure S5. Full NMR spectra (stacked graph) of the complex of BMIE-Cys with zinc chloride.

Each graph includes five spectra, increasing in molar equivalents of the zinc salt from 0 to 2.5 from the bottom upward.



**Figure 46-5.** Figure S6. Full NMR spectra (stacked graph) of the complex of BMIE-Cys with zinc triflate.

Each graph comprises five spectra, increasing in molar equivalents of the zinc salt from 0 to 2.5 from the bottom upward.



**Figure 47-5.** Figure S7. Full NMR spectra (stacked graph) of the complex of BMIE-Cys with tetrakis acetonitrile copper (I) tetrafluoroborate.

Each graph includes five spectra, increasing in molar equivalents of the copper salt from 0 to 2.5 from the bottom upward.

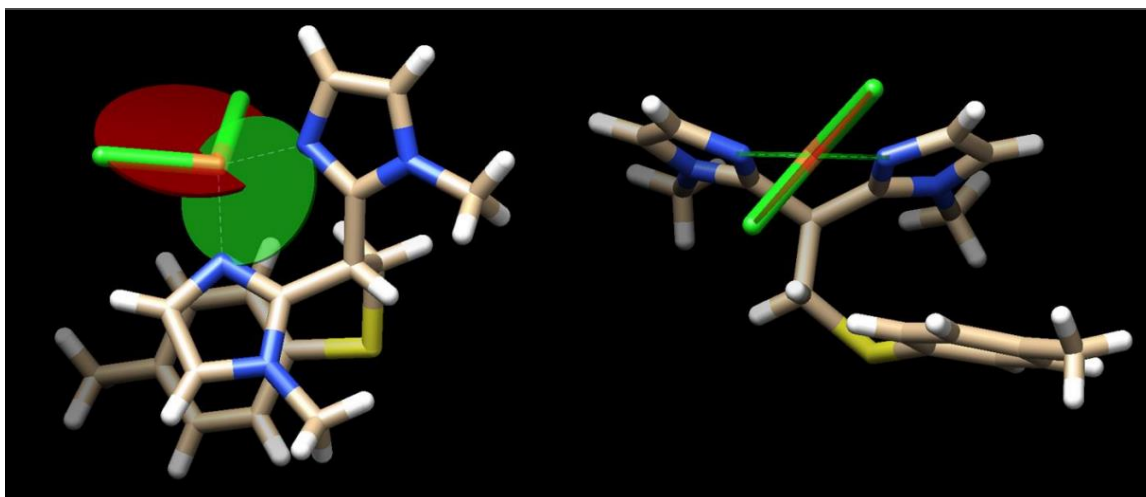
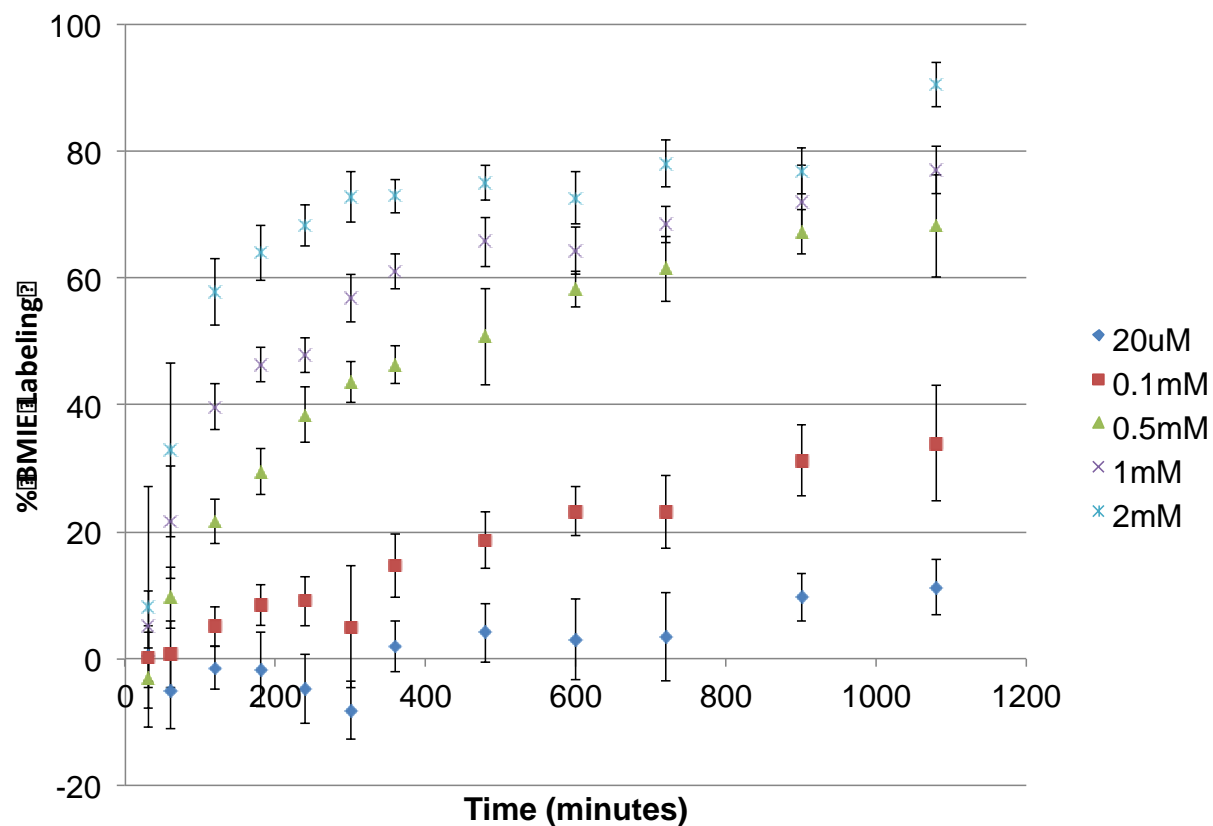


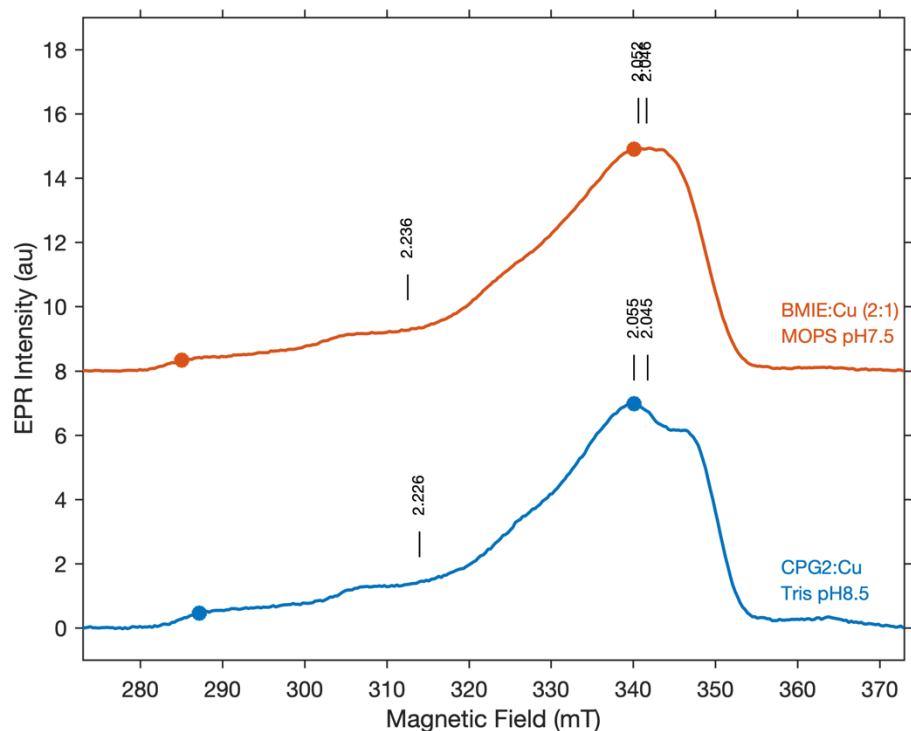
Figure 48-5. Figure S8. Crystal structure of copper chloride in complex with the BMIE-TC ligand.

To highlight the distorted tetrahedral coordination geometry, the plane made by atoms N—Cu—N is represented by a green circle. The plane made by the atoms Cl—Cu—Cl is represented by a red circle. Two different views are shown.



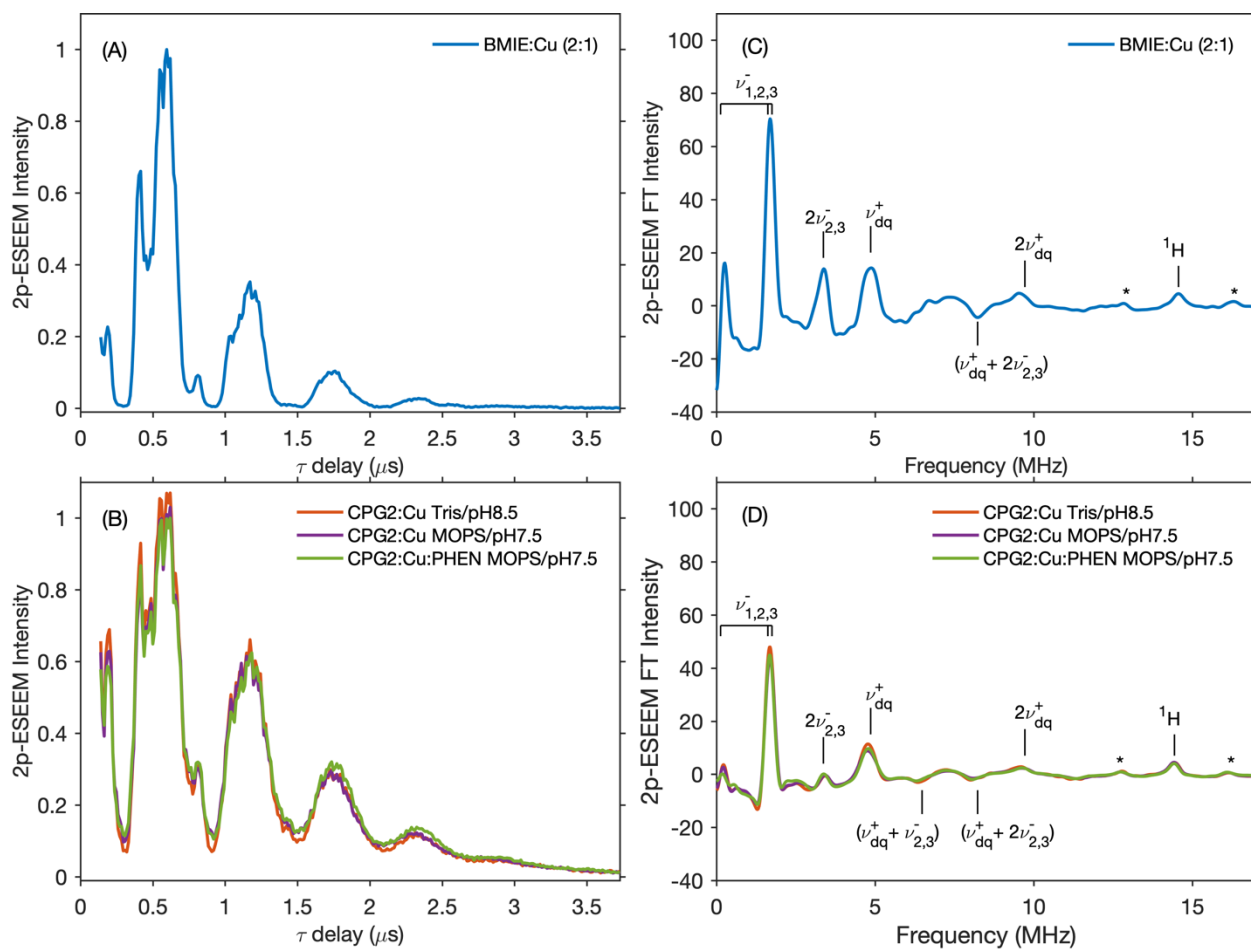
**Figure 49-5.** Figure S9. Time series Ellman's assay of 20uM protein in 100mM NaPO<sub>4</sub>, 1mM EDTA, pH 8.0 with varying concentrations of BMIE.

Aliquots were taken in triplicate at each measurement and reacted with Ellman's reagent to measure the amount of free cysteine. Here we are plotting the amount the  $1 - [\text{free-cys}]/[\text{total protein}]$ .



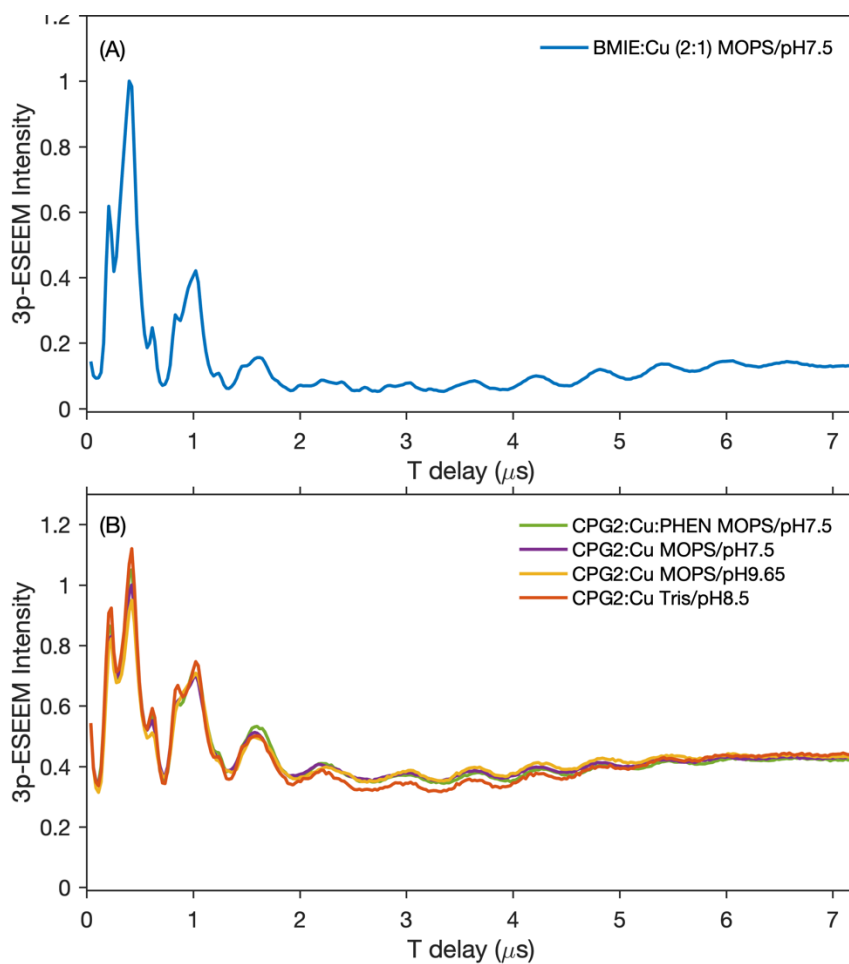
**Figure 50-5.** Figure S10. Two-pulse echo-detected EPR spectra for a small-molecule complex BMIE:Cu (2:1) in MOPS buffer at pH 7.5 (top) and a binary CPG2-BMIE:Cu cluster in Tris buffer at pH 8.5 (bottom), measured at 22 K.

The interpulse delay  $\tau = 552$  ns was used in both experiments. Vertical lines mark the characteristic g-factor values as determined from continuous wave EPR simulations of the same samples (see Fig. X(A) and Table X in the main text). Solid circles indicate two field positions where orientation-selective pulsed EPR experiments were performed (e.g. Fig. X in the main text and Figs. S2-S7 in Supp Info). The field positions 287 mT and 340 mT (solid circles) correspond to the  $g_{\parallel\parallel}$  and  $g_{\perp\perp}$  field orientations, respectively.

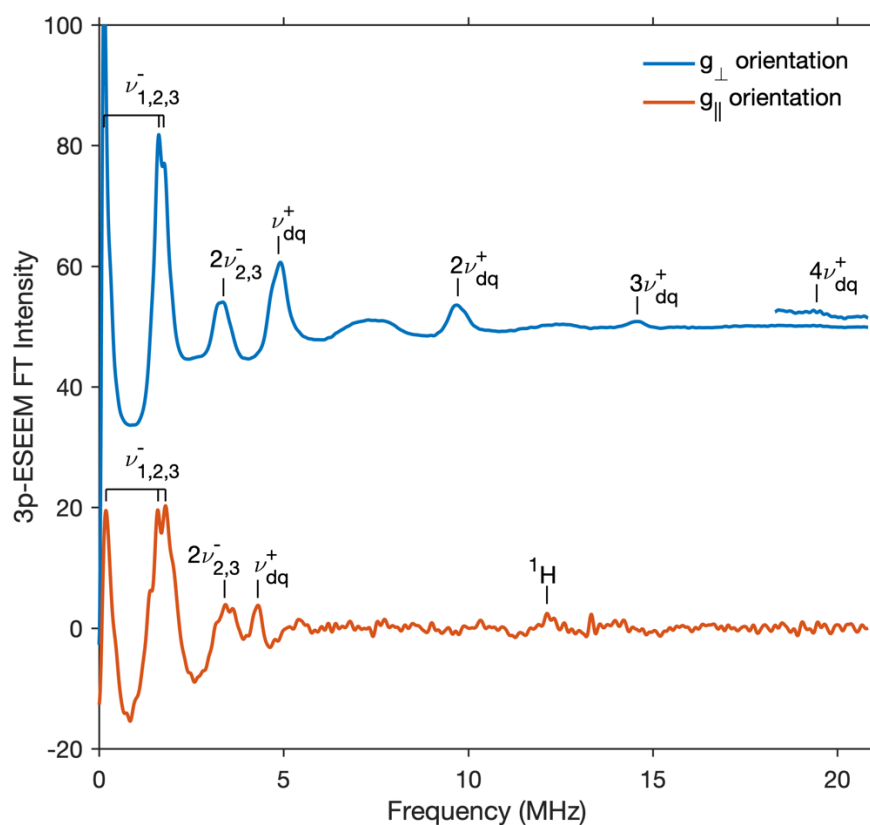


**Figure S11.** Two-pulse ESEEM decays (A, B) and their cosine Fourier-transform (FT) spectra (C, D) for a small-molecule complex BMIE:Cu (2:1) in MOPS buffer at pH 7.5 and binary/ternary CPG2-BMIE:Cu clusters in Tris and MOPS buffers at pH 7.5-8.5, as labeled.

The spectra were measured at 22 K and magnetic field 338-340 mT (the  $g_{\parallel}$  field orientations). The cosine FT spectra were phase corrected for initial delay  $\tau = 140$  ns. The basic  $^{14}N$  nuclear spin transitions ( $\nu_i^-$  and  $\nu_{dq}^+$ ), their harmonics ( $2\nu_{2,3}^-$  and  $2\nu_{dq}^+$ ) and combination peaks ( $\nu_{dq}^+ + n\nu_{2,3}^-$ ) for the amine N-CH<sub>3</sub> nitrogens of Cu-coordinated BMIE ligands are labeled. The stars (\*) mark the positions of the proton-nitrogen combination peaks.

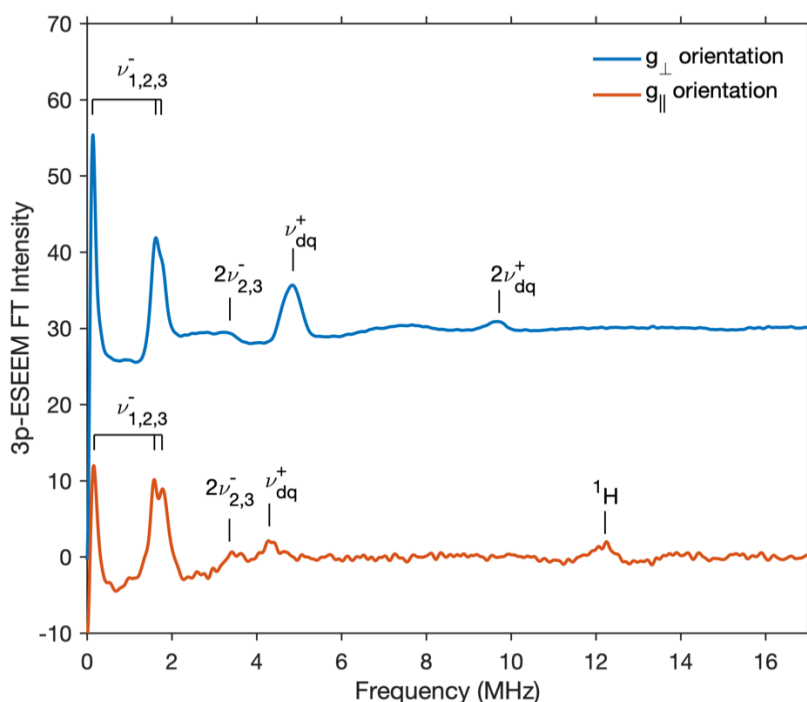


**Figure 52-5.** Figure S12. Three-pulse ESEEM decays for (A) a small-molecule complex BMIE:Cu (2:1) in MOPS buffer at pH 7.5 and (B) binary/ternary CPG2-BMIE:Cu clusters in Tris and MOPS buffers at pH 7.5-9.6, as labeled. The respective Fourier-transform (FT) spectra are shown in Fig. X(B) of the main text. The experiments were performed at 22 K and magnetic field 338-340 mT corresponding to the  $g_{\parallel}$  field orientations. The  $\Delta$  delay between the first two pulses in the three-pulse sequence was set to 204 ns in order to suppress nuclear modulation from weakly-coupled  $^1\text{H}$  protons.



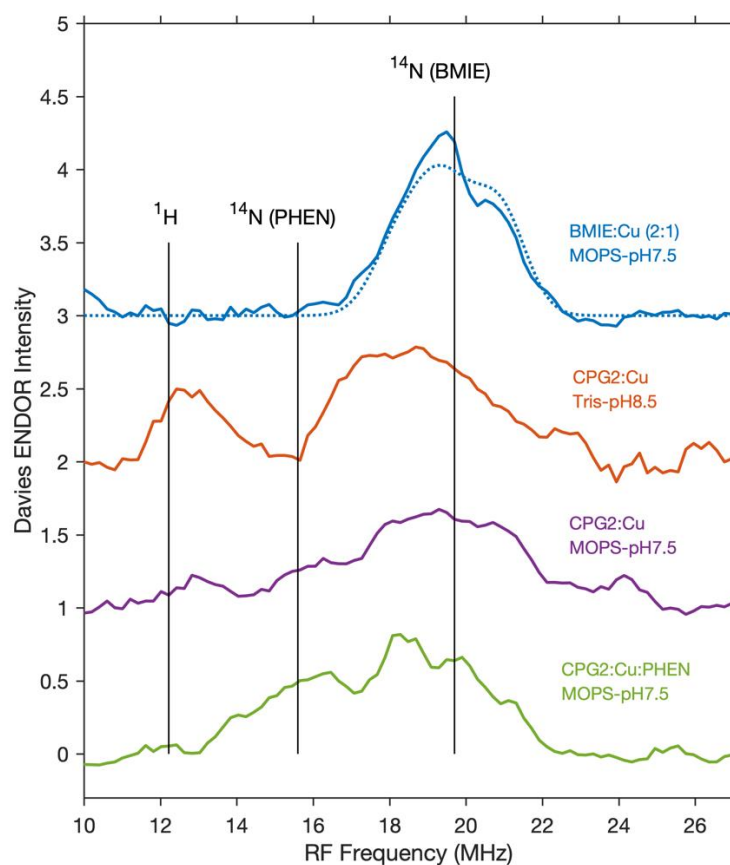
**Figure S3-5.** Figure S13. Cosine FT spectra of three-pulse ESEEM experiment for a small-molecule complex BMIE:Cu (2:1) in MOPS buffer at pH 7.5.

The experiments were performed at 22 K and magnetic fields 340 mT (top) and 285 mT (bottom) corresponding to the  $g_{\perp}$  and  $g_{\parallel}$  field orientations, respectively. The  $\tau$  delays between first two pulses in a three-pulse sequence were set to 204 ns (top) and 140 ns (bottom). The cosine FT spectra were phase corrected to account for the  $\tau$  delay and also for the initial delay  $T = 40$  ns between second and third pulses. The basic  $^{14}\text{N}$  nuclear spin transitions ( $\nu_i^-$  and  $\nu_{dq}^+$ ) and their harmonics ( $2 \cdot \nu_{2,3}^-$  and  $n \cdot \nu_{dq}^+$ ) for amine N-CH<sub>3</sub> nitrogens of Cu-coordinated BMIE are labeled. Four harmonics  $n \cdot \nu_{dq}^+$  ( $n=1-4$ ) are resolved in the ESEEM spectrum at the  $g_{\perp}$  field orientation, indicating four(!) equivalent amine CH<sub>3</sub>-nitrogens (two symmetric BMIE ligands) being coordinated to Cu<sup>2+</sup> in the BMIE:Cu (2:1) complex.



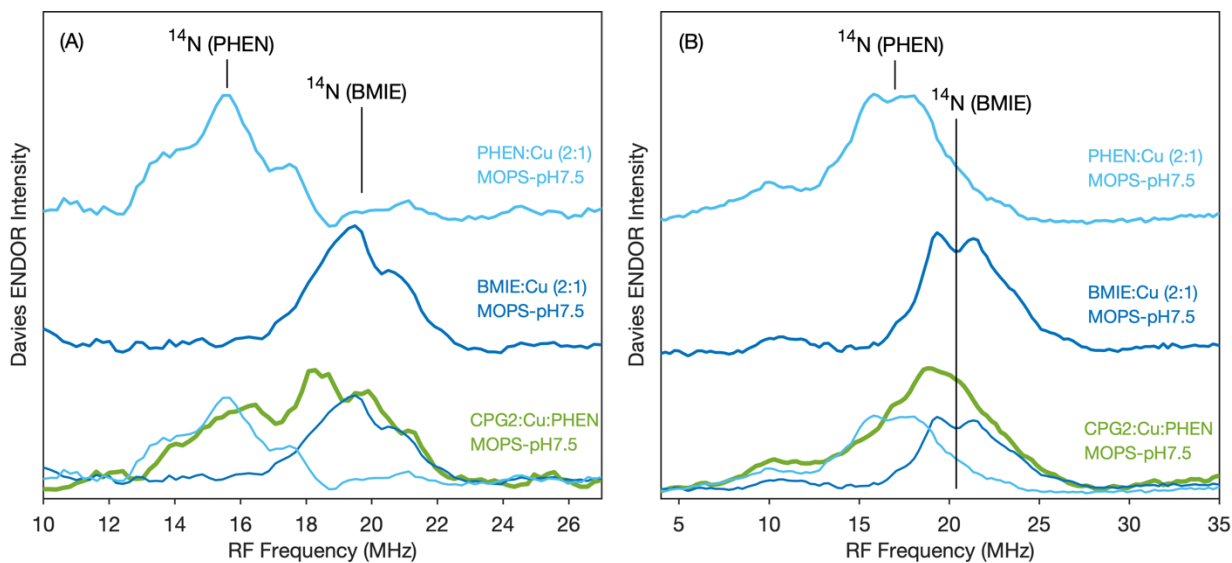
**Figure 54-5.** Figure S14. Cosine FT spectra of three-pulse ESEEM experiment for a binary complex CPG2-BMIE:Cu in Tris buffer at pH 8.5.

The experiments were performed at 22 K and magnetic fields 340 mT (top) and 287 mT (bottom) corresponding to the  $g_{\perp}$  and  $g_{\parallel}$  field orientations, respectively. The  $\tau$  delays between first two pulses in a three-pulse sequence were set to 204 ns (top) and 180 ns (bottom). The cosine FT spectra were phase corrected to account for the  $\tau$  delay and for the initial delay  $T = 40$  ns between second and third pulses. The basic  $^{14}\text{N}$  nuclear spin transitions ( $\nu_i^-$  and  $\nu_{dq}^+$ ) and their harmonics ( $n \cdot \nu_{dq}^+$ ) for amine N-CH<sub>3</sub> nitrogen of BMIE ligand coordinated to Cu<sup>2+</sup> are labeled. Only two harmonics  $n \cdot \nu_{dq}^+$  ( $n=1-2$ ) are observed at the  $g_{\perp}$  field orientation, indicating two(!) equivalent amine CH<sub>3</sub>-nitrogens (one BMIE ligand) coordinated to Cu<sup>2+</sup> in the CPG2-BMIE:Cu complex. Consistently, the nitrogen peak intensities are about a factor of 2 weaker in the CPG2-BMIE:Cu spectra as compared to the BMIE:Cu (2:1) spectra (Fig. S4). Notice the different vertical scales in Fig. S4 and S5.



**Figure S15.** Davies ENDOR spectra for a small-molecule complex BMIE:Cu (2:1) in MOPS buffer at pH 7.5 and several binary/ternary CPG2-BMIE:Cu clusters in Tris and MOPS buffers at pH 7.5-8.5, as labeled, measured at 22 K and magnetic field 285-287 mT corresponding to the  $g_{\parallel}$  field orientations.

The Davies ENDOR spectra measured at the  $g_{\parallel}$  field orientations for the same samples are shown in the main text in Fig. X(C). Non-selective microwave pulses (□ pulse = 32 ns) were applied in order to suppress contributions from weakly-coupled  $^1\text{H}$  protons. Vertical lines mark the peak positions from the BMIE- and PHEN-derived  $^{14}\text{N}$  nitrogens directly-coordinated to  $\text{Cu}^{2+}$ . The dashed line for BMIE:Cu (2:1) shows the simulated ENDOR spectrum using parameters from Table X.



**Figure S16.** Davies ENDOR spectra for two small-molecule complexes BMIE:Cu (2:1) and PHEN:Cu (2:1) and a ternary CPG2-BMIE:Cu:PHEN cluster.

Samples in in MOPS buffer at pH 7.5, as labeled, measured at 22 K and magnetic fields (A) 285 mT and (B) 337-342 mT corresponding to the  $g_{\parallel}$  and  $g_{\perp}$  field orientations, respectively. Non-selective microwave pulses (pulse = 32 ns) were applied in order to suppress contributions from weakly-coupled  $^1\text{H}$  protons. Vertical lines mark the peak positions from the BMIE- and PHEN-derived  $^{14}\text{N}$  nitrogens directly-coordinated to  $\text{Cu}^{2+}$ . On both panels the bottom thin line traces show that the spectra in CPG2-BMIE:Cu:PHEN (green) can be roughly represented as a 1:1 sum of the spectra in BMIE:Cu (2:1) (dark blue) and PHEN:Cu (2:1) (light blue) at both  $g_{\parallel}$  and  $g_{\perp}$  field orientations.

## Chapter 6: Design and Evolution of a Macrocyclic Peptide Inhibitor of the Sonic Hedgehog/Patched Interaction

### 6.1 Preface

A version of this chapter has been published in the *Journal of Chemistry and Science USA* and is formatted in the journal style.

### 6.2 Abstract

The hedgehog (Hh) signaling pathway plays a central role during embryonic development, and its aberrant activation has been implicated in the development and progression of several human cancers. Significant efforts toward the identification of chemical modulators of the hedgehog pathway have yielded several antagonists of the GPCR-like smoothened receptor. In contrast, potent inhibitors of the sonic hedgehog/patched interaction, the most upstream event in ligand-induced activation of this signaling pathway, have been elusive. To address this elusive target, a genetically encoded cyclic peptide was designed based on the sonic hedgehog (Shh)-binding loop of hedgehog-interacting protein (HHIP) and subjected to multiple rounds of affinity maturation through the screening of macrocyclic peptide libraries produced in *E. coli* cells. Using this approach, an optimized macrocyclic peptide inhibitor (HL2-m5) was obtained that binds Shh with a  $K_D$  of 170 nM, which corresponds to a 120-fold affinity improvement compared to the parent molecule. Importantly, HL2-m5 can effectively suppress Shh-mediated hedgehog signaling and Gli-controlled gene transcription in living cells ( $IC_{50} = 230$  nM), providing the most potent inhibitor of the sonic hedgehog/patched interaction reported to date. This first-in-class macrocyclic peptide modulator of the

hedgehog pathway is expected to provide a valuable probe for investigating and targeting ligand-dependent hedgehog pathway activation in cancer and other pathologies. This work also introduces a general strategy for the development of cyclopeptide inhibitors of protein–protein interactions.

### 6.3 Introduction

The hedgehog (Hh) signaling pathway plays a central role during embryonic development, controlling cell growth and differentiation, tissue patterning, and organogenesis.<sup>(1)</sup> Stimulation of the hedgehog pathway is mediated by a complex sequence of molecular events at the level of the membrane and primary cilia of vertebrate cells, resulting in an intracellular signaling cascade and transcriptional response ([Figure 1](#)).<sup>(2)</sup> Canonical activation of this pathway is initiated by binding of the hedgehog signaling proteins (i.e., sonic (Shh), Indian (Ihh), and/or desert (Dhh) hedgehog) to the extracellular domain of the transmembrane receptor patched (PTCH1).<sup>(3)</sup> This event relieves patched-mediated inhibition on the smoothened (Smo) receptor, allowing Smo to translocate from the plasma membrane and endoplasmic vesicles to the primary cilium.<sup>(4)</sup> Smo activation results in the accumulation of the active forms of Gli2 and Gli3 transcription factors,<sup>(5)</sup> which stimulate the transcription of Gli-controlled genes, including *Gli1* and *PTCH1* ([Figure 1](#)).<sup>(6)</sup>

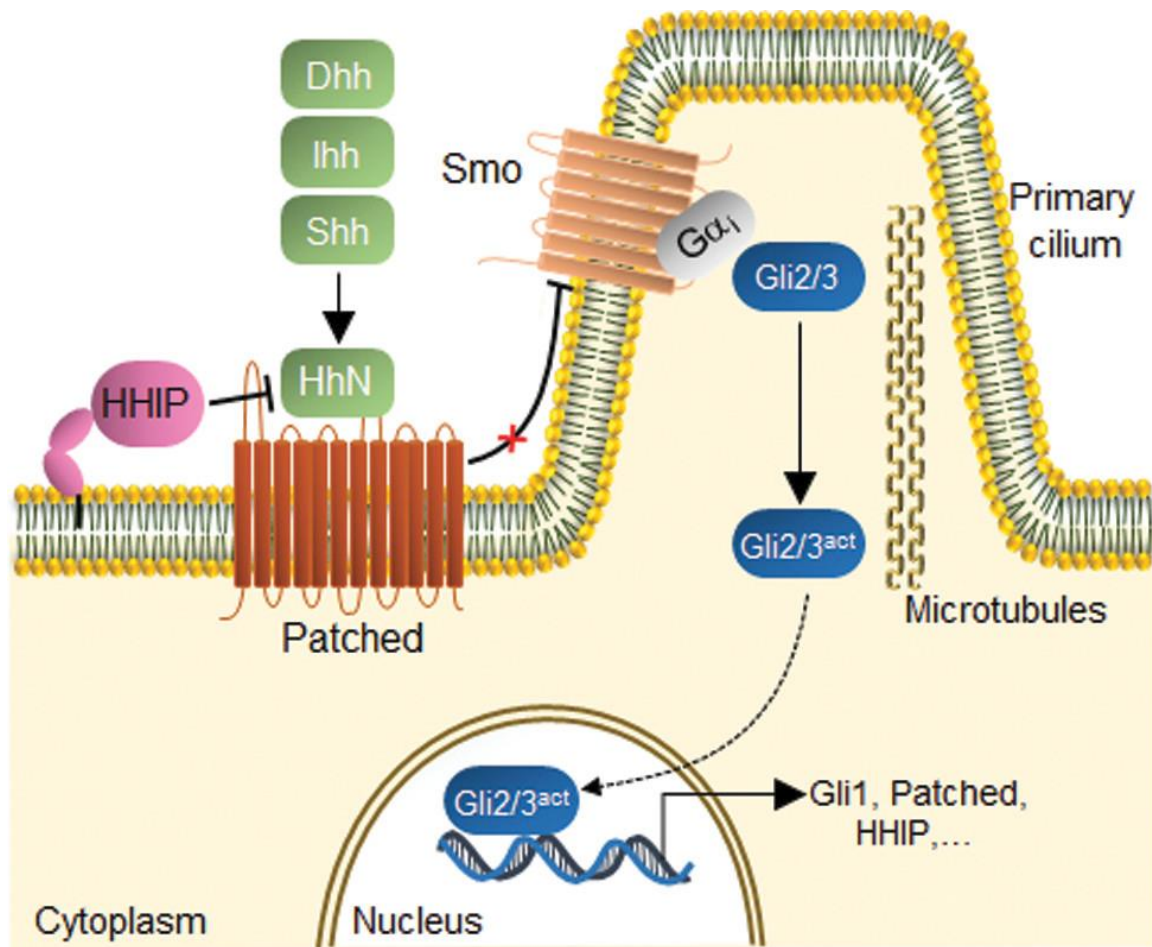


Figure 57-6. Figure 1. Hedgehog signaling pathway.

Binding of the hedgehog ligand(s) (HhN, corresponding to Shh, Dhh, or Ihh) to the patched receptor relieves its inhibitory effect on smoothened (Smo), resulting in the activation of Gli transcription factors and induction of Gli-controlled genes. Hedgehog-interacting protein (HHIP) inhibits the signaling pathway by competing with patched for binding to the hedgehog ligands. Adapted from [www.phosphosite.org](http://www.phosphosite.org).<sup>20</sup>

Aberrant activation of the hedgehog pathway has been associated with tumorigenesis in several human tissues.<sup>(7)</sup> In particular, an increasing number of tumors, including leukemia,<sup>(8)</sup> small-cell lung,<sup>(9)</sup> pancreas,<sup>(10)</sup> and colon<sup>(11)</sup> cancer, have been found to rely on ligand-dependent hedgehog signaling for sustained growth and proliferation. Hh signaling is also implicated in the maintenance and propagation of cancer stem cells,<sup>(8a,</sup>

[8c, 12](#)) which are believed to play a crucial role in tumor self-renewal, survival against chemotherapy, and metastasis.[\(13\)](#)

Because of the therapeutic potential of the hedgehog pathway, significant efforts have been devoted to the development of chemical modulators of this signaling pathway.[\(14\)](#) These efforts have led to the identification of several potent inhibitors of GPCR-like smoothened receptor.[\(14\)](#) These include cyclopamine[\(15\)](#) and vismodegib,[\(16\)](#) which correspond to the archetypal member and the first FDA-approved drug, respectively, belonging to this class of Hh pathway antagonists. Compounds that target downstream components of this pathway[\(14\)](#) or processes involved in Shh maturation[\(17\)](#) have also been reported. In stark contrast, potent inhibitors of the Shh/patched protein–protein interaction have remained elusive. While a neutralizing anti-hedgehog antibody (5E1) is available,[\(18\)](#) small-molecular-weight agents capable of disrupting this interaction would be desirable. To date, the only compound of this type is robotnikinin, a small-molecule Shh antagonist developed by Schreiber and co-workers.[\(19\)](#) Despite this progress, this compound has only moderate Shh inhibitory activity *in vitro* and *in cellulo* ( $IC_{50} \approx 15 \mu M$ ),[\(19\)](#) highlighting the need for more potent inhibitors directed against this component of the hedgehog pathway.

Macrocyclic peptides are promising molecular scaffolds for targeting biomolecular interfaces, including those mediating protein–protein interactions.[\(21\)](#) Because of their attractive features as chemical probes and potential therapeutics, we previously developed methodologies to access macrocyclic peptides through the cyclization of ribosomally derived polypeptides using a genetically encoded noncanonical amino acid (ncAA).[\(22\)](#) ncAA-mediated peptide cyclization offers the opportunity to rapidly

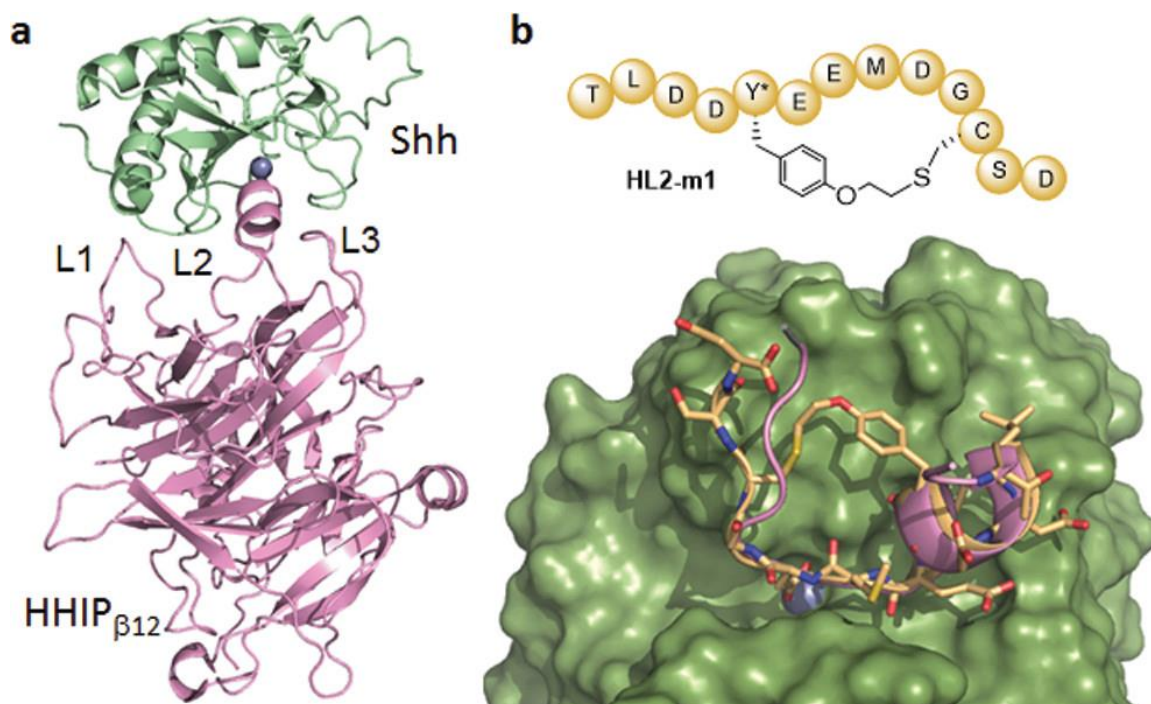
generate genetically encoded cyclic peptide libraries directly in bacterial cells, which can facilitate molecular discovery efforts.<sup>(23)</sup> Here, we successfully applied this strategy to develop and evolve a macrocyclic peptide that targets Shh with high binding affinity and effectively inhibits Shh-mediated hedgehog pathway signaling in living mammalian cells. This work makes available a valuable probe molecule for investigating the functional role and therapeutic potential of the Shh/patched interaction. In addition, it introduces and validates an integrated platform for the development of bioactive macrocyclic peptides.

## 6.4 Results and Discussion

### 6.4.1 Design of Shh-Binding Macrocyclic Peptide HL2-m1

Recent crystallographic studies have provided insights into the structure of Shh in complex with hedgehog-interacting protein (HHIP), a membrane protein that acts as a negative regulator of the Hh pathway ([Figure 1](#)).<sup>(24)</sup> In this complex, HHIP is found to interact with Shh primarily via an extended loop (L2) located in the extracellular domain of HHIP ([Figure 2a](#)).<sup>(25)</sup> These previous studies also indicated that the Shh-binding site involved in the interaction with the HHIP L2 loop is shared by patched, as evinced by (a) the sequence similarity between HHIP L2 and an L2-like sequence within the patched extracellular domain and (b) the ability of a linear HHIP L2-derived peptide (HHIP<sub>370–390</sub>) to inhibit the Shh/patched interaction *in vitro*, albeit with only very weak activity (IC<sub>50</sub>: 150  $\mu$ M).<sup>(25a)</sup> Based on this information, we envisioned that a macrocyclic peptide encompassing the HHIP L2 loop sequence would provide a viable starting point for the development of an agent capable of disrupting the Shh/patched protein–protein interaction. In particular, we recognized that the distance between the  $\alpha$ -carbon atoms of residue Met379 and Leu385 within the L2 loop of HHIP ([Figure S1](#)) is compatible with

the interside-chain thioether bridge provided by a peptide macrocyclization method previously reported by our group.[\(22a\)](#) The latter involves a cross-linking reaction between a cysteine residue and a genetically encodable *O*-2-bromoethyltyrosine (O2beY), which bears a cysteine-reactive alkyl bromide group.[\(22a\)](#) The side chains of the Met379 and Leu385 residues point away from the HHIP L2-binding cleft in Shh, suggesting that a bridge connecting these positions would not directly interfere with Shh binding. At the same time, the conformational restriction imposed by the interside-chain linkage was expected to be beneficial toward improving Shh affinity compared to a linear L2-derived peptide, as a result of reduced entropic costs upon binding to the protein. Based on these considerations, a molecular model of the resulting O2beY/Cys-bridged peptide, called HL2-m1, was generated and docked into the structure of Shh using Rosetta simulations. Briefly, viable conformations of HL2-m1 that accommodate the thioether cross-link was generated based on the crystal structure of the HHIP-Shh complex, followed by energy minimization using the Rosetta FastRelax protocol[\(26\)](#) in the modeled Shh-bound state. These analyses provided support to the design by showing a good overlap between the backbone of the modeled cyclic peptide and that of the HHIP L2 loop in the Shh-bound structure as well as the absence of steric clashes between the thioether bridge and the L2-binding cleft in the Shh protein ([Figure 2b](#)).



**Figure 58-6.** Figure 2. Macrocyclic HHIP L2 loop mimic.

(a) Crystal structure of Shh (green) in complex with the extracellular domain of HHIP (pink) (PDB 3HO5(25a)). The three loop regions of HHIP involved in Shh binding are labeled and the zinc ion in the L2-binding cleft of Shh is shown as a sphere model (blue). (b) Top: schematic structure of the macrocyclic peptide HL2-m1. Bottom: the model of HL2-m1 (yellow, stick model) bound to Shh (green, surface model). The L2 loop of HHIP (pink, ribbon model) is superimposed to the modeled complex.

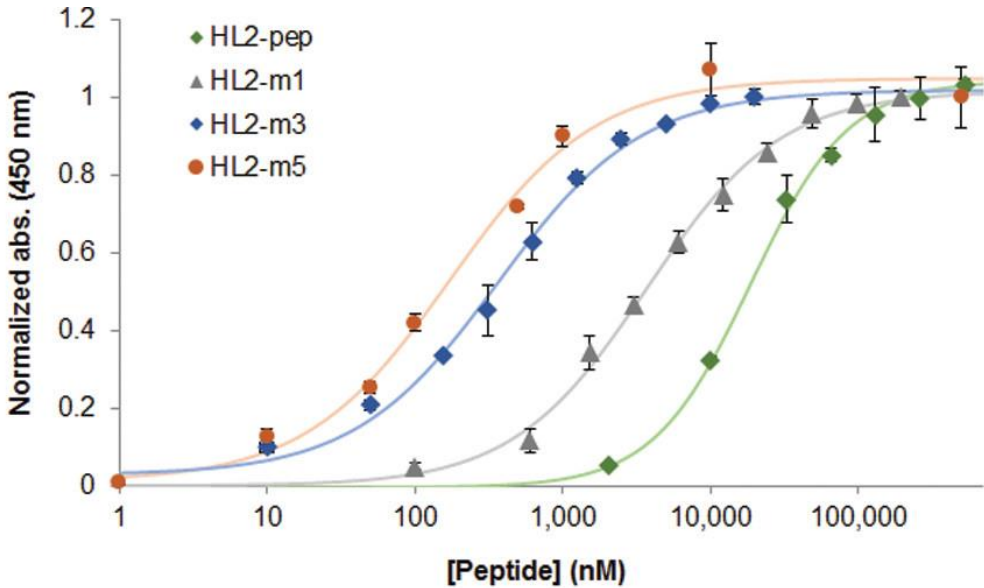
#### 6.4.2 Characterization of HL2-m1

To assess both the biosynthetic accessibility and Shh-binding properties of HL2-m1, the designer cyclic peptide was targeted for production by recombinant means. To this end, a gene encoding for a 13mer peptide sequence spanning the HHIP L2 loop (HHIP<sub>375–387</sub>) was cloned into a pET-based expression vector. The codon corresponding to the Met5 residue in the HL2-m1 sequence (Met375 in HHIP) was mutated to an amber stop codon (TAG) to allow for the site-selective incorporation of O2beY via amber stop codon

suppression.<sup>(27)</sup> Residue Leu11 (Leu385 in HHIP) was mutated to cysteine to mediate the formation of the desired thioether bridge. The distance between these residues ( $i/i + 6$ ) was expected to be compatible with O2beY/Cys cyclization based on our previous studies with model peptide sequences.<sup>(22a)</sup> The HL2-m1 encoding sequence was then fused to an N-terminal FLAG tag for detection purposes (*vide infra*) and to a C-terminal GyrA intein<sup>(28)</sup> containing a polyhistidine tag to facilitate purification and isolation. The resulting polypeptide construct (FLAG-HL2m1-T-GyrA-H<sub>6</sub>) was expressed in *E. coli* cells in the presence of O2beY and an O2beY-specific orthogonal aminoacyl-tRNA synthetase/tRNA pair.<sup>(22a)</sup> After protein purification via Ni-affinity chromatography, the FLAG-tagged HL2-m1 peptide was cleaved from the GyrA intein with thiophenol, followed by HPLC purification. MALDI-TOF MS analysis of the thiol-induced cleavage reaction mixture showed the release of the desired cyclic peptide and no detectable amounts of the acyclic peptide (Figure S6a), indicating that O2beY/Cys cyclization had occurred efficiently and quantitatively upon expression in *E. coli* cells.

To measure the Shh-binding affinity of HL2-m1, an *in vitro* assay was developed in which recombinant GST-fused Shh was immobilized on microtiter plates and then exposed to the FLAG-tagged peptide. The Shh-bound peptide is then quantified colorimetrically ( $\lambda_{450}$ ) using horseradish peroxidase (HRP)-conjugated anti-FLAG antibody. Using this assay, the FLAG-HL2-m1 peptide was determined to bind Shh with a  $K_D$  of 3.6  $\mu$ M (Figure 3). In comparison, a linear peptide encompassing the L2 sequence (FLAG-L2-pep) exhibited a significantly lower binding affinity for Shh ( $K_D = 20 \mu$ M). These results demonstrated the functionality of the designer macrocyclic L2 loop mimic as an Shh targeting agent. In addition, the 5.5-fold higher Shh binding affinity of HL2-m1

compared to its linear counterpart highlighted the anticipated beneficial effect of macrocyclization toward stabilizing the bioactive conformation of the peptide.



Peptide	Sequence	$K_D$ (nM)
FLAG-HL2-pep	X- <sup>1</sup> TLDDMEEMDGLS <sup>13</sup> DT	$20,000 \pm 1,000$
FLAG-HL2-m1	X-TLDD(O2beY)EEMDGCSDT	$3,600 \pm 200$
FLAG-HL2-m3	X-TLDW(O2beY)EEMDMCTDT	$330 \pm 30$
FLAG-HL2-m5	X-TLSW(O2beY)EAMDMCTDT	$170 \pm 20$

Figure 59-6. Figure 3. The shh-binding affinity of linear and macrocyclic L2 mimics.

(a) Dose–response curves for direct binding of the recombinantly produced FLAG tag-fused peptides to plate-immobilized GST-Shh as determined using HRP-conjugated anti-FLAG antibody. (b) Sequences and  $K_D$  values corresponding to macrocyclic L2 mimics and linear L2-based peptide. X = MDYKDDDDKGGSGS-. The mutated positions in the evolved macrocyclic peptides compared to the initial cyclic peptide HL2-m1 are highlighted.

### 6.4.3 Affinity Maturation of Macrocyclic HHIP L2 Loop Mimics

Next, we sought to improve the Shh binding affinity of HL2-m1 by leveraging the ability to encode and produce the macrocyclic peptide in bacterial cells genetically. To this end, the strategy outlined in [Figure 4](#) was applied, which entails the generation of HL2-m1 variant libraries in multiwell plates followed by screening of the recombinantly produced macrocyclic peptides directly in cell lysates. In this system, the amino acid residue (Thr) at the junction between the macrocycle-encoding sequence and the GyrA intein was removed to leave an Asp residue at the "intein-1" position. This residue was previously determined to promote the spontaneous, post-translational cleavage of the C-terminal intein moiety directly in cells, thereby releasing the peptide macrocycle.[\(22a\)](#) Upon cell lysis, the FLAG-tagged macrocycles are screened for improved Shh binding activity using the colorimetric assay described above.

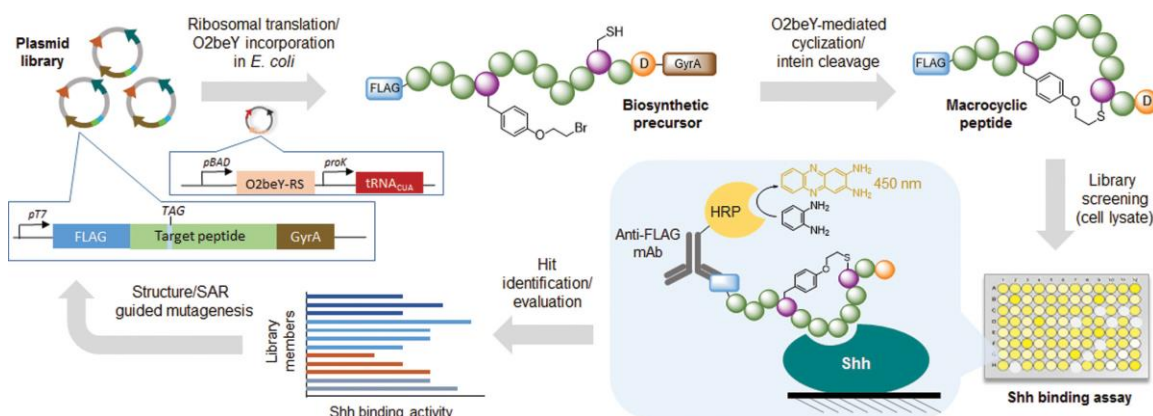


Figure 60-6. Figure 4. Overview of strategy for the evolution of macrocyclic peptides.

A plasmid library encoding for partially randomized peptide sequences fused to a FLAG tag and a C-terminal GyrA intein is transformed into *E. coli* cells and arrayed on multiwell plates. The corresponding precursor polypeptides are produced via ribosomal translation and O2beY incorporation via amber stop codon (TAG) suppression. The macrocyclic

peptides are produced inside cells through "self-processing" of the biosynthetic precursors via O2beY/Cys cyclization and aspartate-induced intein cleavage. After cell lysis, peptide binding to immobilized Shh is quantified colorimetrically. The variants showing improved Shh binding activity are deconvoluted via DNA sequencing. The best variant and acquired SAR data are used for the next round of affinity maturation.

For the first round of affinity maturation, five positions within the HL2-m1 macrocycle were selected for mutagenesis based on the modeled HL2-m1/Shh complex ([Figure 2b](#)). These positions include three interfacial residues located within the  $\alpha$ -helical (Glu6, Glu7) and turn region (Gly10) of the molecule and two solvent-exposed residues neighboring the O2beY/Cys linkage (Asp4, Ser12). In this and subsequent steps (*vide infra*), residue Asp9 was left unaltered since the corresponding residue in HHIP (Asp383) mediates an energetically important interaction by coordinating a Zn(II) ion in the L2 binding cleft of Shh ([Figure S1](#)).<sup>(25)</sup> Accordingly, five macrocycle libraries were prepared via site-saturation mutagenesis (NNK degenerate codon) of the aforementioned positions within the HL2-m1 encoding sequence. For each library, ~90 recombinant clones were arrayed on 96-well plates, followed by in-cell production of the corresponding macrocycles. Upon screening of the library with the immunoassay, several library members were found to exhibit improved Shh-binding activity compared to the parent compound HL2-m1 ([Figure S3](#)). Interestingly, each of the single-site libraries yielded two or more improved variants. Among them, the variant containing a Ser12Met mutation, renamed **HL2-m2**, emerged as the most promising hit, and it was thus selected as the reference compound during the second round of affinity maturation.

Importantly, detailed structure–activity information for each mutated site was gathered at this point by sequencing the multiple hits identified from the initial macrocycle libraries

([Figure S3](#)). Based on this information, three second-generation libraries were prepared by recombining beneficial mutations at positions 4 (A/D/G/W), 6 (L/S/V/W/E), 7 (K/Y/A/E), 10 (G/M/T), and 12 (L/M/T/S). Upon screening of the resulting libraries (~500 recombinants) according to the strategy of [Figure 4](#), a macrocycle variant (HL2-m3) showing improved Shh-binding activity compared to HL2-m2 was identified ([Figure S4](#)). By sequencing, **HL2-m3** was determined to contain a total of three mutations, namely, Asp4Trp, Gly10Met, and Ser12Thr ([Figure 3](#)). The cyclic structure of HL2-m3 was further confirmed by expressing this sequence as stable GyrA intein fusion (i.e., by introducing Thr at the "intein-1" position), followed by thiol-induced intein cleavage and MS analysis. These tests showed the occurrence of the macrocyclic peptide as the only detectable species ([Figure S6b](#)). After purification, this compound was determined to bind Shh with a  $K_D$  of 330 nM ([Figure 3](#)), corresponding to an 11-fold improvement compared to HL2-m1.

As the next step, all unmodified positions within the **HL2-m3** sequence (relative to HL2-m1) were randomized by site-saturation mutagenesis. From the resulting libraries, an improved HL2-m3-derived variant was obtained, which carries a Glu → Ala mutation at the level of residue 7 (**HL2-m4**). A second hit carrying an Asp3Ser mutation was also identified at this stage. Upon combining these mutations, a further improved Shh-binding macrocycle was obtained, which was named **HL2-m5**. HL2-m5 contains a total of five amino acid substitutions compared to HL1-m1, and it undergoes quantitative cyclization *in vivo* ([Figure S6c](#)), further demonstrating the robustness and reliability of the O2beY-mediated peptide cyclization chemistry. Upon purification, FLAG-HL2-m5 was found to bind Shh with a  $K_D$  of 170 nM ([Figure 3](#)), which corresponds to a more than 20-fold

increase in affinity compared to the initially designed macrocycle (HL2-m1) and a nearly 120-fold improvement compared to the linear L2-derived peptide. Altogether, these results supported the effectiveness of the strategy outlined in [Figure 4](#) toward enabling the affinity maturation of the initially designed macrocyclic L2 mimic.

#### 6.4.4 Molecular Modeling and Circular Dichroism Experiments

To gain further insights into the role of the beneficial mutations accumulated in HL2-m5, a model of the cyclic peptide in complex with Shh was generated using Rosetta simulations ([Figure S2](#)). Inspection of the complex suggested the occurrence of potential interactions between the Trp4 and Met10 residues of HL2-m5 with regions of the Shh surface that are not contacted by the corresponding residues in the HL2-m1 peptide ([Figure 2](#)) or within the L2 loop of HHIP ([Figure S1](#)). Specifically, the side chain of Met10 inserts into the L2 binding cleft of Shh ([Figure S2](#)), establishing new contacts between the protein and the HL2-m5 peptide that are not present in the Shh/HHIP complex due to the presence of a Gly residue at this position. Particularly interesting is also the case of Trp4, whose aryl ring inserts into a nearby cleft on the Shh surface according to the energy-minimized model of the complex. Experimentally, the energetic importance of this interaction is corroborated by the identification of an identical mutation, i.e., Asp4 → Trp, among the most active compounds isolated from the single-site mutagenesis libraries derived from HL2-m1 ([Figure S3](#)). On the other hand, the positive effect of the Asp7Ala substitution accumulated in HL2-m5 is supported by the approximately 2-fold higher Shh binding affinity of HL2-m4 compared to HL2-m3. The same substitution was also identified as beneficial during the screening of the HL2-m1-

derived libraries ([Figure S3](#)). Residue 7 is located at the C-terminal end of the two-turn  $\alpha$ -helix, and the beneficial effect of the alanine substitution at this position can be rationalized based on the stabilization of an  $\alpha$ -helical conformation in this region of the molecule.

To better examine the conformational properties of HL2-m5, a FLAG tag-free version of this peptide along with that of the linear L2-derived peptide (Ac-TLDDMEEMDGLSD-NH<sub>2</sub>) was prepared by solid-phase peptide synthesis (*vide infra*) and analyzed by circular dichroism (CD). As shown in [Figure 5a](#), the near-UV CD spectrum of the linear L2-based peptide is consistent with that of a random coil polypeptide, indicating that it lacks a well-defined structure in solution. In contrast, the HL2-m5 macrocycle exhibits more pronounced negative bands in the 215–222 nm range along with a positive band in the 190–195 nm region of the CD spectrum ([Figure 5b](#)). These features are consistent with the presence of a more structured peptide containing an  $\alpha$ -helical motif. In addition, unlike for the linear peptide, intensification of the spectral features of the HL2-m5 macrocycle was observed upon the addition of the helix-inducing solvent trifluoroethanol.<sup>(29)</sup> Thus, in addition to more favorable contacts with the Shh surface as suggested by molecular modeling, the improved Shh-binding affinity of HL2-m5 compared to the linear L2 peptide likely arises from a stabilization of the bioactive conformation as a result of the cyclic backbone and other sequence alterations (e.g., Asp7Ala mutation).

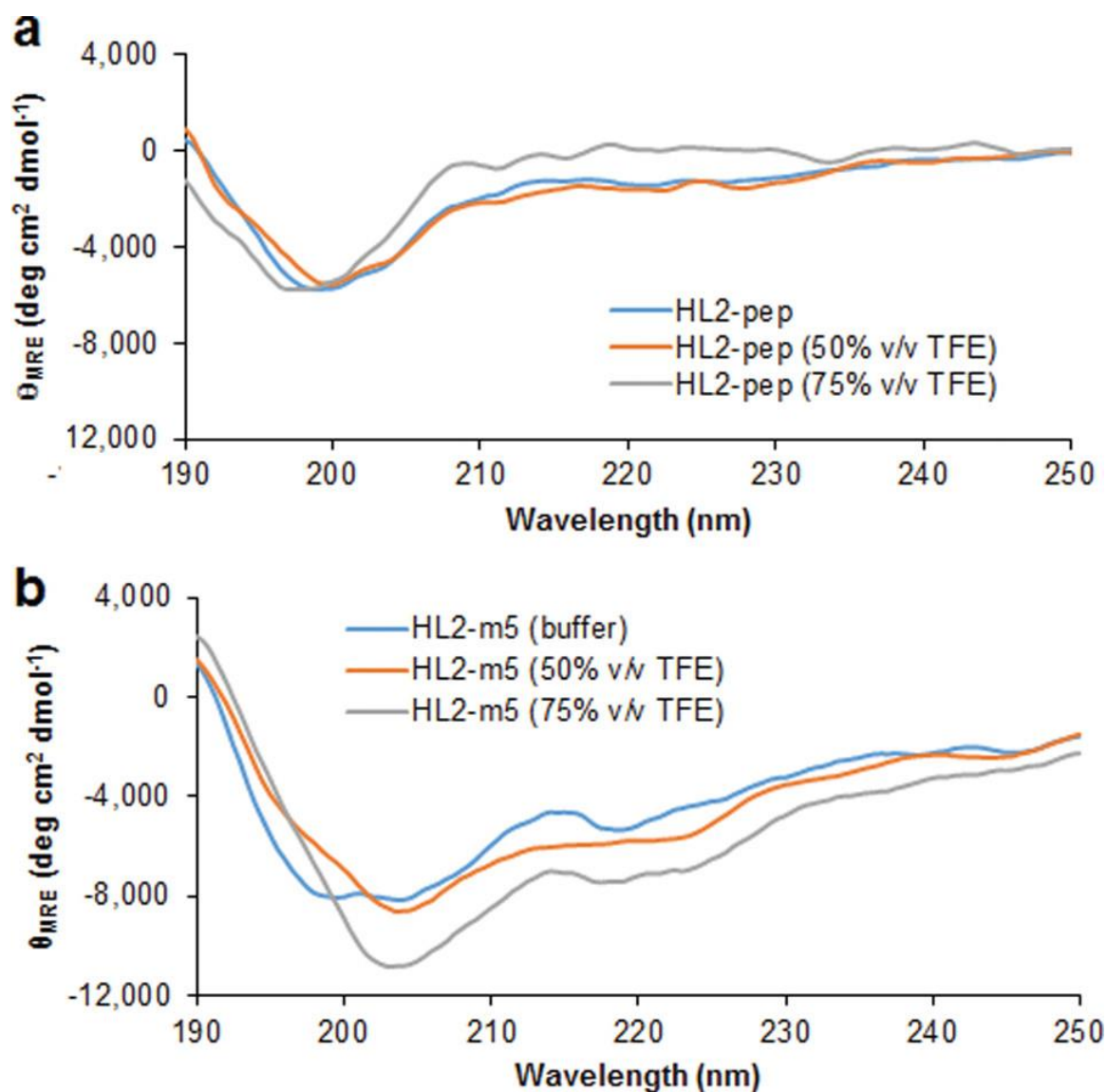


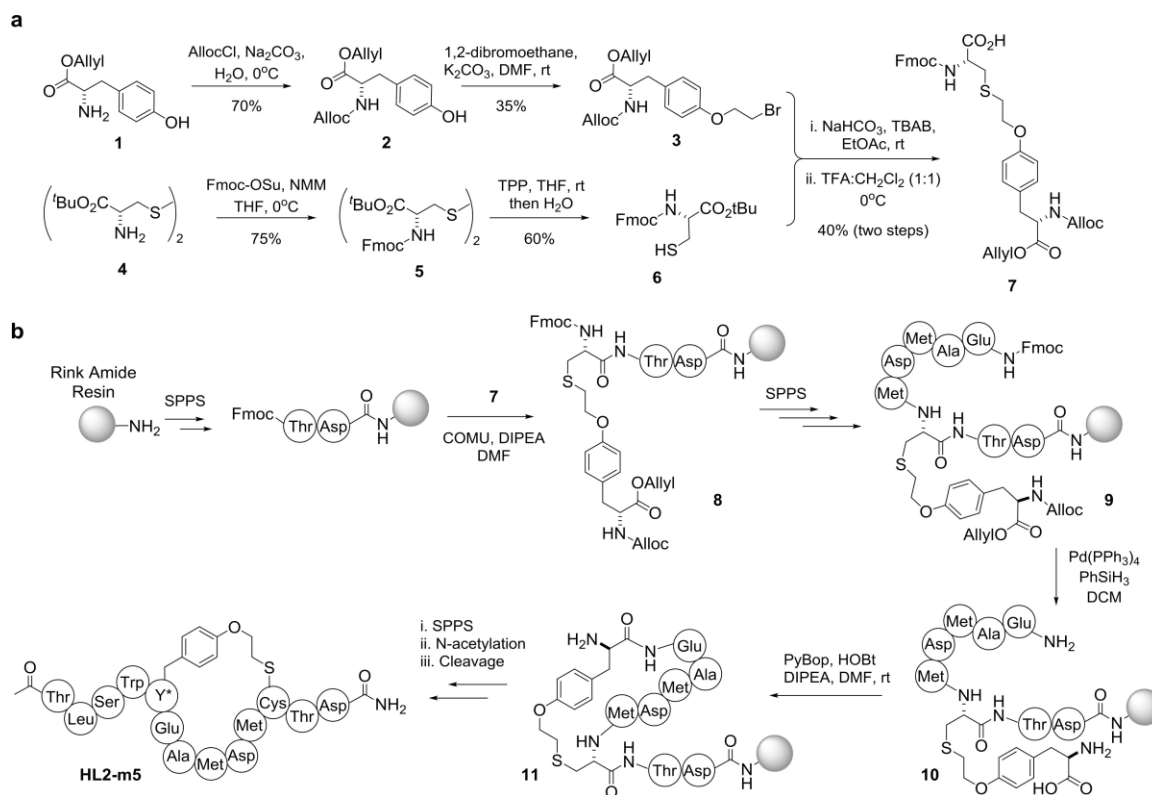
Figure 61-6. Figure 5. Circular dichroism (CD)

Spectra corresponding to the linear HL2-pep (a) and the macrocyclic HL2-m5 peptide (b) in buffer and in the presence of trifluoroethanol (TFE) at varying concentrations. The signals are reported as mean residue molar ellipticity ( $\theta_{MRE}$ ).

#### 6.4.5 Synthesis of Macrocyclic Peptides via SPPS

To provide further access to the macrocyclic peptides, a synthetic strategy was devised to afford these compounds. Inspired by approaches previously adopted for the synthesis of

lantibiotics,<sup>(30)</sup> this strategy involves the incorporation of a dipeptide building block encompassing the O2beY/Cys thioether cross-link during solid-phase peptide synthesis (SPPS), followed by on-resin cyclization and cleavage/deprotection of the peptide from the resin ([Scheme 1](#)). As shown in [Scheme 1a](#), the dipeptide building block **7** was prepared via alkylation of *N*-Alloc-(*O*-2-bromoethyl)tyrosine allyl ester (**3**) with *N*-Fmoc-(1)-cysteine *tert*-butyl ester (**6**), followed by removal of the *tert*-butyl group under acidic conditions. For the synthesis of HL2-m5, the first two C-terminal amino acids were loaded on a Rink amide MBHA resin, followed by incorporation of the dipeptide building block via amide coupling with COMU, yielding **8**. The peptide chain was then further extended to include amino acid residues Met10 to Glu7, affording **9**. The side-chain Alloc and allyl ester protecting groups were then removed using Pd(PPh<sub>3</sub>)<sub>4</sub> catalyst in the presence of PhSiH<sub>3</sub>, whereas the N-terminal amino group was exposed via Fmoc deprotection. On-resin cyclization was then realized under amide coupling conditions with PyBOP and HOBt in the presence of DIPEA, to afford **11**. The peptide was then further extended via SPPS to include the N-terminal tail of the peptide, followed by Fmoc deprotection and N-acetylation. The synthesis of HL2-m5 was completed by cleavage of the peptide from the resin using a 95:2.5:2.5 trifluoroacetic acid/triisopropylsilane/water mixture. After purification by reverse-phase HPLC, the desired macrocyclic peptide was obtained with an overall yield of 15% ([Figure S8](#)). The same protocol could then be applied to afford HL2-m1 ([Figure S9](#)) in comparable yields.



**Figure 62-6.** Scheme 1. Synthesis of Macrocyclic Peptides:

(a) Synthetic Route for the Preparation of the Diamino Acid Building Block Encompassing the Cys/O2beY Thioether Linkage; (b) Solid-Phase Synthesis of Macrocyclic Peptide HL2-m5.

#### 6.4.6 Suppression of Hedgehog Pathway Activation in Living Cells

Having demonstrated the ability of HL2-m5 to target Shh *in vitro*, we next examined its activity toward disrupting Shh-mediated hedgehog pathway signaling in cells. To this end, we utilized a cell-based luciferase reporter assay,<sup>(31)</sup> in which mouse embryo fibroblasts (NIH3T3) are transfected with vectors encoding for firefly luciferase (FF) gene under a Gli-controlled promoter and a *Renilla* luciferase (Ren) gene under a constitutive promoter. Hh pathway suppression is measured based on the decrease in the firefly/*Renilla* luminescence ratio in the presence of the inhibitor. In preliminary

experiments, this assay was validated using the Smo inhibitor cyclopamine, which caused full inhibition of Shh-induced luminescence in the cells at a concentration of 10  $\mu$ M, in accordance with previous reports.[\(31\)](#)

After transfection with the luciferase reporter plasmids, NIH3T3 cells showed strong luminescence in the presence of recombinant *N*-palmitoylated sonic hedgehog (Shh-N) and low luminescence in the absence of Shh-N, thereby confirming Shh-dependent activation of the hedgehog pathway in the cells. Upon incubation of Shh-N-stimulated cells with HL-m5, dose-dependent suppression of the luminescence signal was observed ([Figure 6a](#)), from which a half-maximal inhibitory concentration ( $IC_{50}$ ) of 250 nM was determined. A residual pathway activity was observed at the highest dose tested, which could be attributed to the limited solubility of the peptide at a concentration of  $\geq 5$ –10  $\mu$ M in the medium used for the cell-based assay. In contrast to the cyclic peptide, the linear L2-pep peptide showed no inhibitory activity at concentrations up to 30  $\mu$ M under identical conditions. Incubation of HL-m5-treated cells with purmorphamine, a Smo agonist,[\(32\)](#) restored activation of the signaling pathway ([Figure 6b](#)), demonstrating that HL2-m5-dependent inhibition occurs at the level of Shh/patched interaction. Notably, the inhibitory activity of HL2-m5 toward blocking hedgehog pathway activation in cells is nearly two orders of magnitude higher than that of robotnikinin ( $IC_{50} \approx 15$   $\mu$ M),[\(19\)](#) as determined using a similar cell-based assay. Noteworthy is also the fact that the  $IC_{50}$  value exhibited by HL2-m5 in the cell-based assay (250 nM) is very similar to the  $K_D$  value measured for FLAG-HL2-m5 in the *in vitro* Shh binding assay (170 nM, [Figure 3](#)) and to the  $IC_{50}$  value of HL2-m5 determined using this assay in a competition format (280 nM, [Figure S10](#)). These results indicate that the macrocyclic peptide targets Shh

with high affinity and specificity even in the presence of cells and a complex growth medium.

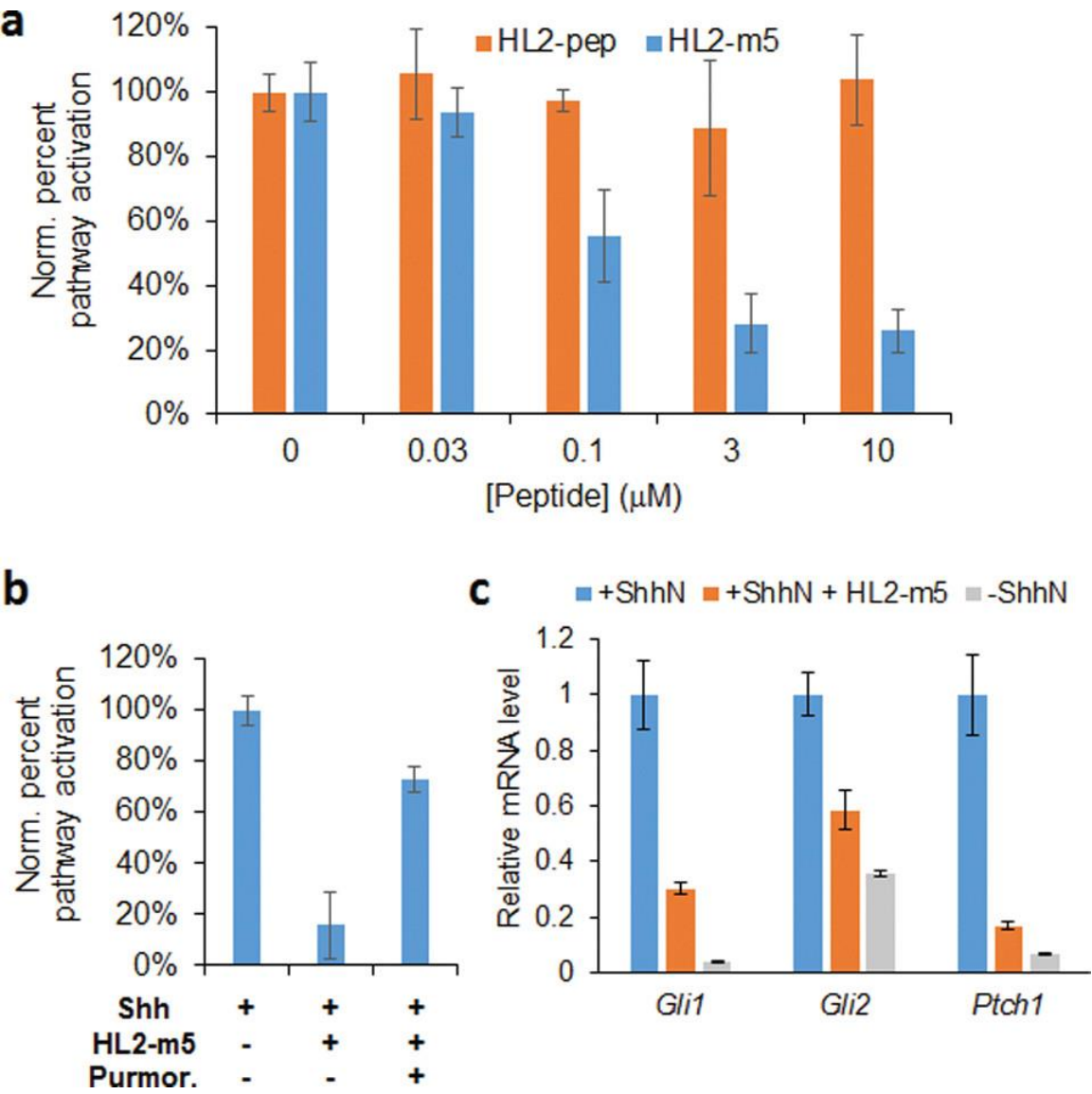


Figure 63-6. Figure 6. HL2-m5-induced suppression of Hh pathway signaling.

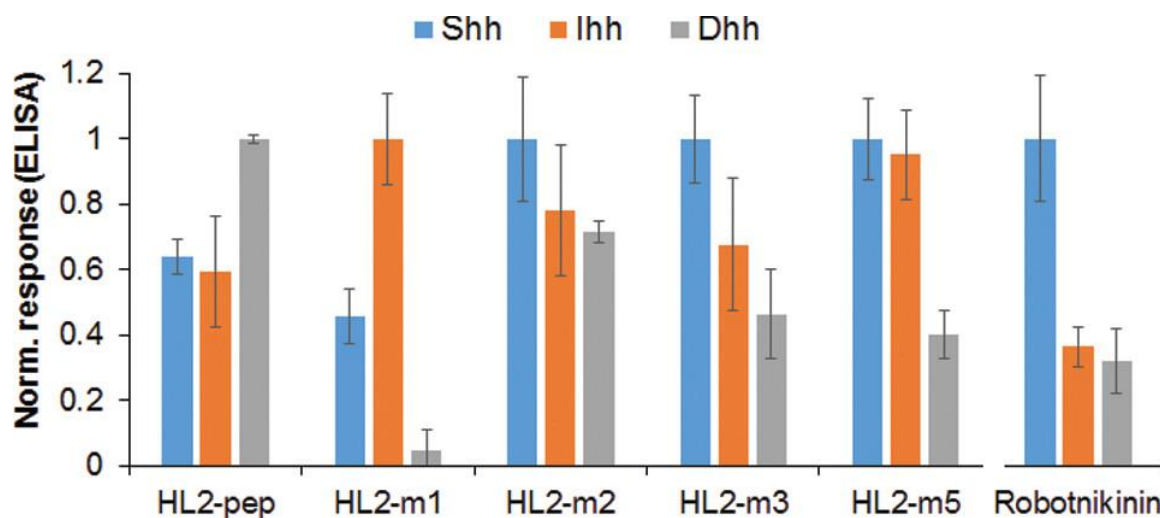
(a) Dose-dependent inhibition of luciferase expression (FF/Ren ratio) in Shh-stimulated NIH3T3 cells containing a dual luciferase reporter system. (b) Restoration of hedgehog pathway signaling upon the addition of purmorphamine (5 μM) to cells treated with HL2-m5 (10 μM). (c) Relative transcriptional levels of *Gli1*, *Gli2*, and *Ptch1* genes in Shh-stimulated NIH3T3 cells in the presence and in the absence of HL2-m5 (10 μM) as determined by real-time PCR. mRNA levels in unstimulated cells are included for comparison.

To further validate HL2-m5 as a hedgehog pathway antagonist, the effect of this compound on the transcriptional activity of two canonical target genes of the pathway, *Gli1* and *Ptch1* ([Figure 1](#)), was examined via real-time PCR. As shown in [Figure 6c](#), a significant reduction (75–85%) of the mRNA levels corresponding to these genes was observed in Shh-N-stimulated cells upon incubation with HL2-m5 at 10  $\mu$ M, relative to compound-untreated cells. Treatment with the macrocyclic peptide also suppresses the mRNA level for the transcription factor Gli2. For both *Ptch1* and *Gli2*, the corresponding transcriptional levels in HL2-m5-treated cells approach those observed in unstimulated cells grown in the absence of Shh-N ligand ([Figure 6c](#)). No changes in cell morphology, growth behavior, and titer were noted in the presence of HL2-m5, indicating a lack of cytotoxicity at the highest concentration range applied in these experiments. Taken together, these results demonstrate that the macrocyclic peptide can potently inhibit Shh-dependent hedgehog pathway activation in living cells and suppress signature transcriptional responses resulting from ligand-induced stimulation of the pathway.

#### 6.4.7 Hedgehog Analog Selectivity

While Shh is the most abundant analog among hedgehog proteins, paracrine/autocrine hedgehog signaling in normal and cancer cells is also mediated by the Indian (Ihh) and Desert (Dhh) analogs.[\(33\)](#) Hh-targeted inhibitors capable of targeting multiple analogs of this protein are thus expected to be particularly useful toward suppressing ligand-induced activation of this pathway. Since the Hh analog selectivity of robotnikinin had not previously been investigated, this property was examined using a competition assay, whereby inhibition of FLAG-HL2-m5 binding to plate-immobilized GST-Shh, GST-Ihh,

or GST-Dhh is measured via the HRP-conjugated anti-FLAG antibody. These experiments showed that robotnikinin has a significantly lower affinity toward Ihh and Dhh relative to Shh ([Figure 7](#)). By comparison, HL2-m5 was found to interact with all three analogs of Hh proteins, showing nearly identical activity toward Shh and Ihh ([Figure 7](#)). Consistent with this trend, direct binding experiments showed that HL2-m5 interacts with Ihh and Dhh with a  $K_D$  of 160 and 330 nM, respectively ([Figure S11](#)). Thus, the affinity of HL2-m5 for Ihh and Dhh is nearly identical and only 2-fold lower, respectively, than that for Shh (170 nM) as determined using the same assay. These results indicate that the macrocyclic peptide can act as an effective inhibitor for all analogs of the hedgehog protein.



**Figure 64-6.** Figure 7. Hh analog selectivity of linear and cyclic L2 mimics.

Data relative to the peptides are derived from direct binding experiments to immobilized Hh proteins. Data relative to robotnikinin are derived from competition experiments (10  $\mu$ M robotnikinin + 400 nM FLAG-HL2-m5). For each compound, values are normalized to the highest binding response measured across the three hedgehog analogs.

In the interest of determining how the affinity maturation process affected the Hh analog selectivity of the peptides, these experiments were extended to the other linear and cyclic L2 mimics. These analyses showed that the linear L2-based peptide (HL2-pep) binds preferentially to Dhh over Shh and Ihh ( $\text{Dhh} > \text{Shh} \approx \text{Ihh}$ ; [Figure 7](#)). This selectivity profile mirrors that of full-length HHIP.<sup>(34)</sup> Interestingly, macrocyclization of the L2 peptide sequence via the O2beY/Cys linkage (= HL2-m1) results in a complete shift in Hh analog selectivity ( $\text{Ihh} > \text{Shh} \gg \text{Dhh}$ ), leading to a preference for Ihh over Shh and nearly abolishing its affinity for Dhh. In the HL2-m1  $\rightarrow$  HL2-m2 transition, the selectivity is then shifted toward Shh. This result is reasonable given that improved Shh binding was the selection criteria applied during the library screening process. At the same time, it is interesting to see how a single mutation accumulated during this step (Ser12Met) restores binding to Dhh and leads to Hh analog cross-reactivity ([Figure 7](#)). With HL-m3, the preference for Shh over the other two Hh analogs becomes more pronounced ( $\text{Shh} > \text{Ihh} \approx \text{Dhh}$ ). As noted above, HL2-m5 shows comparable affinity toward Shh and Ihh and higher preference toward these analogs over Dhh ( $\text{Shh} \approx \text{Ihh} > \text{Dhh}$ ). Altogether, these results illustrate the potential of tuning the Hh analog selectivity of these macrocyclic peptide scaffolds.

## 6.5 Conclusion

In summary, we have reported the development of a potent macrocyclic peptide inhibitor of the Shh/patched interaction, an essential protein–protein interaction implicated in the activation of the hedgehog pathway. HL2-m5 binds Shh with high affinity *in vitro* and can effectively suppress Shh-mediated stimulation of hedgehog pathway signaling in

living mammalian cells. The inhibitory activity of HL2-m5 is about two orders of magnitude higher than that of robonitnikinin, the only compound previously reported to target the Shh/patched interaction. Furthermore, unlike robonitnikinin; HL2-m5 exhibits a high affinity toward all three analogs of the hedgehog protein. HL2-m5 also shows promising stability against proteolytic degradation ( $t_{1/2} > 6\text{--}8$  h in blood serum; [Figure S12](#)). Collectively, these features should make HL2-m5 a valuable probe for investigating the biological role and therapeutic potential of the Hh/patched interaction in the context of pathologies that are associated with aberrant ligand-dependent activation of the hedgehog pathway.

From a methodological standpoint, this work demonstrates the value of the strategy outlined in [Figure 4](#) toward the development of potent and selective macrocyclic peptide disruptors of protein–protein interactions. Using this approach, a low-affinity linear peptide encompassing a Shh recognition motif from HHIP could be rapidly evolved into a high-affinity Shh-targeting agent (120-fold lower  $K_D$ ) through the generation and screening of macrocyclic peptide libraries generated in bacteria. This process was further facilitated by the ability to produce and isolate the macrocyclic peptides by recombinant means, which expedites hit evaluation in secondary functional assays. At the same time, an efficient methodology was implemented to afford these compounds by synthetic means, which will facilitate further optimization of these compounds using non-proteinogenic amino acids. We expect that the overall strategy presented here will prove valuable for the development of bioactive cyclopeptides against a variety of other challenging protein–protein interactions.

## 6.6 References

1.
  - a. Chiang C, Litlington Y, Lee E, Young KE, Corden JL, Westphal H, and Beachy PA. **Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function.** *Nature* 1996, **383**:407-13.
  - b. Ingham PW, McMahon AP. **Hedgehog signaling in animal development: paradigms and principles.** *Genes & development* 2001, **23**:3059-87.
  - c. Ingham PW, Placzek M: **Orchestrating ontogenesis: variations on a theme by sonic hedgehog.** *Nature Reviews Genetics* 2006, **11**:841-50.
2.
  - a. Ingham PW, McMahon AP: **Hedgehog signaling in animal development: paradigms and principles.** *Genes & development* 2001, **23**:3059-87.
  - b. Jiang J, Hui CC: **Hedgehog signaling in development and cancer.** *Developmental cell* 2008, **15**:801-12.
  - c. Wong SY, Reiter JF: **The primary cilium: at the crossroads of mammalian hedgehog signaling.** *Current topics in developmental biology* 2008, **85**:225-60.
3. Carpenter LR, Farruggella TJ, Symes A, Karow ML, Yancopoulos GD, Stahl N: **Enhancing leptin response by preventing SH2-containing phosphatase 2 interaction with Ob receptor.** *Proceedings of the National Academy of Sciences* 1998, **95**: 6061-6.

4. Deneff N, Neubüser D, Perez L, Cohen SM. **Hedgehog induces opposite changes in turnover and subcellular localization of patched and smoothened.** *Cell* 2000, **102**:521-31.
5.
  - a. Wen X, Lai CK, Evangelista M, Hongo JA, de Sauvage FJ, Scales SJ: **Kinetics of hedgehog-dependent full-length Gli3 accumulation in primary cilia and subsequent degradation.** *Molecular and cellular biology* 2010, **30**:1910-22.
  - b. Kim J, Kato M, Beachy PA: **Gli2 trafficking links Hedgehog-dependent activation of Smoothened in the primary cilium to transcriptional activation in the nucleus.** *Proceedings of the National Academy of Sciences* 2009, **106**:21666-71.
6.
  - a. Yoon JW, Kita Y, Frank DJ, Majewski RR, Konicek BA, Nobrega MA, Jacob H, Walterhouse D, Iannaccone P: **Gene expression profiling leads to identification of GLI1-binding elements in target genes and a role for multiple downstream pathways in GLI1-induced cell transformation.** *Journal of Biological Chemistry* 2002, **277**:5548-55.
  - b. Kasper M, Regl G, Frischauf AM, Aberger F: **GLI transcription factors: mediators of oncogenic Hedgehog signalling.** *European Journal of Cancer*. 2006, **42**:437-45.
- 7.

- a. Rubin LL, de Sauvage FJ: **Targeting the Hedgehog pathway in cancer.** *Nature reviews Drug discovery* 2006, **5**:1026-33.
  - b. Theunissen JW, de Sauvage FJ: **Paracrine Hedgehog signaling in cancer.** *Cancer research* 2009, **69**:6007-10.
- 8.
- a. Dierks C, Beigi R, et al.: **Expansion of Bcr-Abl-Positive Leukemic Stem Cells Is Dependent on Hedgehog Pathway Activation.** *Cancer Cell* 2008, **14**: 238-249.
  - b. Hegde GV, Peterson KJ, Emanuel K, Mittal AK, Joshi AD, Dickinson JD, Kollessery GJ, Bociek RG, Bierman P, Vose JM, Weisenburger DD: **Hedgehog-induced survival of B-cell chronic lymphocytic leukemia cells in a stromal cell microenvironment: a potential new therapeutic target.** *Molecular Cancer Research* 2008, **6**:1928-36.
  - c. Zhao C, Chen A, Jamieson CH, Fereshteh M, Abrahamsson A, Blum J, Kwon HY, Kim J, Chute JP, Rizzieri D, Munchhof M: **Hedgehog signalling is essential for maintenance of cancer stem cells in myeloid leukaemia.** *Nature* 2009, **458**:776-9.
9. Watkins DN, Berman DM, Burkholder SG, Wang B, Beachy PA, Baylin SB. Hedgehog signalling within airway epithelial progenitors and in small-cell lung cancer. *Nature*. 2003 Mar;422(6929):313-7.
10. Thayer, S., di Magliano, M., Heiser, P. et al.: **Hedgehog is an early and late mediator of pancreatic cancer tumorigenesis.** *Nature* 2003, **425**: 851–856.

11. Berman, D., Karhadkar, S., Maitra, A. et al.: **Widespread requirement for Hedgehog ligand stimulation in growth of digestive tract tumours.** *Nature* 2003, **425**: 846–851.
12. Liu S, Dontu G, Mantle ID, Patel S, Ahn NS, Jackson KW, Suri P, Wicha MS: **Hedgehog signaling and Bmi-1 regulate self-renewal of normal and malignant human mammary stem cells.** *Cancer research* 2006, **66**:6063-71.
13.
  - a. Reya T, Morrison SJ. clarke, MF & Weissman, IL: **Stem cells, cancer, and cancer stem cells.** *Nature* 2001, **414**:105-1.
  - b. Beachy PA. Karhadkar SS, Berman DM: **Tissue repair and stem cell renewal in carcinogenesis.** *Nature* 2004, **432**:324-31.
14.
  - a. Stanton BZ, Peng LF: **Small-molecule modulators of the Sonic Hedgehog signaling pathway.** *Molecular BioSystems* 2010, **6**:44-54.
  - b. Peukert S, Miller-Moslin K: **Small-molecule inhibitors of the hedgehog signaling pathway as cancer therapeutics.** *ChemMedChem: Chemistry Enabling Drug Discovery* 2010, **5**:500-12.
  - c. Sharpe HJ, Wang W, Hannoush RN, De Sauvage FJ: **Regulation of the oncoprotein Smoothed by small molecules.** *Nature chemical biology* 2015, **4**:246.
15. Cooper MK. porter JA, young KE, Beachy PA: **Teratogen-mediated inhibition of target tissue response to Shh signaling.** *Science* 1998, **280**:1603-7.

16. Robarge KD, Brunton SA, Castanedo GM, Cui Y, Dina MS, Goldsmith R, Gould SE, Guichert O, Gunzner JL, Halladay J, Jia W: **GDC-0449—a potent inhibitor of the hedgehog pathway.** *Bioorganic & medicinal chemistry letters*, **19**:5576-81.
17. Owen TS, Xie XJ, Laraway B, Ngoje G, Wang C, Callahan BP. **Active site targeting of hedgehog precursor protein with phenylarsine oxide.** *Chembiochem* 2015, **16**:55-8.
18.
  - a. Ericson J, Morton S., et al.: **Two Critical Periods of Sonic Hedgehog Signaling Required for the Specification of Motor Neuron Identity.** *Cell* 1996, **87**: 661-673.
  - b. Beachy PA, Hymowitz SG, Lazarus RA, Leahy DJ, Siebold C: **Interactions between Hedgehog proteins and their binding partners come into view.** *Genes & development* 2010, **24**:2001-12.
19. Stanton BZ, Peng LF, Maloof N, Nakai K, Wang X, Duffner JL, Taveras KM, Hyman JM, Lee SW, Koehler AN, Chen JK. A. Mandinova, and SL Schreiber: **A small molecule that binds Hedgehog and blocks its signaling in human cells.** 2009:154-6.
20. Hornbeck, P. V.; Zhang, B.; Murray, B.; Kornhauser, J. M.; Latham, V.; Skrzypek, E: **PhosphoSitePlus, 2014: mutations, PTMs and recalibrations.** *Nucleic Acids Res.* 2015, **43**: 512-520.
- 21.

- a. Driggers, E., Hale, S., Lee, J. et al. The exploration of macrocycles for drug discovery — an underexploited structural class. *Nat Rev Drug Discov* 7, 608–624 (2008).
- b. Robinson, J.A. et al.: **The design, structures and therapeutic potential of protein epitope mimetics.** *Drug Discovery Today* 2008, **13**: 944-951.
- c. Marsault, E.; Peterson, M: **Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery.** *Med Chem* 2011, **54**: 1961-2004.
- d. Hill TA, Shepherd NE, Diness F, Fairlie DP: **Fixierung cyclischer Peptide: Mimetika von Proteinstrukturmotiven.** *Angewandte Chemie* 2014, **126**:13234-57.
- e. Cardote TA, Ciulli A: **Cyclic and macrocyclic peptides as chemical tools to recognise protein surfaces and probe protein–protein interactions.** *ChemMedChem* 2016, **11**:787-94.
- f. Villar EA, Beglov D, Chennamadhavuni S, Porco Jr JA, Kozakov D, Vajda S, Whitty A: **How proteins bind macrocycles.** *Nature chemical biology* 2014, **10**: 723.

22.

- a. Bionda N, Cryan AL, Fasan R: **Bioinspired strategy for the ribosomal synthesis of thioether-bridged macrocyclic peptides in bacteria.** *ACS chemical biology* 9:2008-13.
- b. Frost JR, Jacob NT, Papa LJ, Owens AE, Fasan R: **Ribosomal synthesis of macrocyclic peptides in vitro and in vivo mediated by genetically**

- encoded aminothiol unnatural amino acids.** *ACS chemical biology* 2015, **10**:1805-16.
- c. Bionda N, Fasan R: **Ribosomal synthesis of natural-product-like bicyclic peptides in Escherichia coli.** *ChemBioChem* 2015, **16**:2011-6.
23. Tavassoli A: **SICLOPPS cyclic peptide libraries in drug discovery.** *Current opinion in chemical biology* 2017, **38**:30-5.
24. Chuang PT, McMahon AP: **Vertebrate Hedgehog signalling modulated by induction of a Hedgehog-binding protein.** *Nature* 1999, **397**:617-21.
- 25.
- a. Bosanac I, Maun HR, Scales SJ, Wen X, Lingel A, Bazan JF, De Sauvage FJ, Hymowitz SG, Lazarus RA: **The structure of SHH in complex with HHIP reveals a recognition role for the Shh pseudo active site in signaling.** *Nature structural & molecular biology* 2009 **16**:691.
- b. Bishop B, Aricescu AR, Harlos K, O'callaghan CA, Jones EY, Siebold C: **Structural insights into hedgehog ligand sequestration by the human hedgehog-interacting protein HHIP.** *Nature structural & molecular biology* 2009, **16**:698.
- 26.
- a. Fleishman SJ, Leaver-Fay A, Corn JE, Khare SD, Koga N: **RosettaScripts: an XMLlike interface to the Rosetta macromolecular modeling suite.** *PLoS One* 2011.
- b. Tyka MD, Keedy DA, André I, DiMaio F, Song Y, Richardson DC, Richardson JS, Baker D: **Alternate states of proteins revealed by**

- detailed energy landscape mapping.** *Journal of molecular biology* 2011, **405**:607-18.
27. Liu, C. C.; Schultz, P. G: **Adding New Chemistries to the Genetic Code.** *Annu Rev Biochem* 2010, **79**:413-444.
  28. Smith, J.M., Vitali, F., Archer, S.A. and Fasan, R: **Modular Assembly of Macrocyclic Organo–Peptide Hybrids Using Synthetic and Genetically Encoded Precursors.** *Angew. Chem. Int. Ed.* 2011, **50**: 5075-5080
  29. Nelson, JW, Kallenbach, NR: **Stabilization of the ribonuclease S-peptide  $\alpha$ -helix by trifluoroethanol.** *Proteins* 1986, **1**: 211-217.
  30.
    - a. Pattabiraman, V., McKinnie, S. and Vederas, J: **Solid-Supported Synthesis and Biological Evaluation of the Lantibiotic Peptide Bis(desmethyl) Lacticin 3147 A2.** *Angewandte Chemie International Edition* 2008, **47**: 9472-9475
    - b. Knerr, P et al.: **Synthesis and Activity of Thioether-Containing Analogs of the Complement Inhibitor Compstatin.** *ACS Chem Bio* 2011, **7**: 753-760.
  31. Chen, J.K. et al.: **Small molecule modulation of Smoothed activity.** *PNAS* 2002.
  32. Sinha, S., Chen, J: **Purmorphamine activates the Hedgehog pathway by targeting Smoothed.** *Nat Chem Biol* 2006, **2**: 29–30.
  - 33.

- a. Azoulay S et al.: **Comparative expression of Hedgehog ligands at different stages of prostate carcinoma progression.** *J. Pathol* 2008, **216**: 460-470.
  - b. Ibuki, N., Ghaffari, M., Pandey, M., Iu, I., Fazli, L., Kashiwagi, M., Tojo, H., Nakanishi, O., Gleave, M.E. and Cox, M.E: **TAK-441, a novel investigational smoothened antagonist, delays castration-resistant progression in prostate cancer by disrupting paracrine hedgehog signaling.** *Int. J. Cancer* 2013, **133**: 1955-1966.
34. Martinez-Chinchilla P and Riobo N.A.: **Purification and Bioassay of Hedgehog Ligands for the Study of Cell Death and Survival.** *Methods in Enzymology* 2008, **446**: 189-206.
- 35.
- a. Zanghellini A et al.: **New algorithms and an in silico benchmark for computational enzyme design.** *Protein Science* 2006, **15**: 2785-2794
  - b. Richeter F. et al.: **De Novo Enzyme Design Using Rosetta3.** *PLOS One* 2011, **6**.
36. Bionda N, Fasan R: **Ribosomal Synthesis of Thioether-Bridged Bicyclic Peptides.** *Methods Mol. Biol.* 2017, 57-76.

## 6.7 Experimental Procedures

### 6.7.1 Cloning, Expression, and Purification of GST-Hh Proteins

Vectors containing human Shh, Ihh, and Dhh genes were kindly provided by the Riobo-Del Galdo laboratory.[\(34\)](#) Genes encoding for Shh, Ihh, and Dhh were amplified by PCR (primers #1–6; [Table S1](#)) and cloned into the *Nco* I/*Xho* I cassette of the expression vector pET42b (Novagen), resulting in the C-terminal fusion of the Hh protein sequence to that of glutathione-S-transferase (GST) protein containing a polyhistidine tag. The GST-Hh fusion proteins were expressed in BL21(DE3) cells by growing recombinant cells in LB medium with kanamycin (30 µg/mL). At an OD<sub>600</sub> of 0.6, cells were induced with 1 mM IPTG (isopropyl-β-D-1-thiogalattopyranoside) and grown for 20 h at 27 °C. The proteins were purified by Ni-NTA chromatography (Invitrogen) according to the manufacturer's instructions. After elution, the proteins were buffer exchanged with phosphate-buffered saline (PBS) buffer (10 mM Na<sub>2</sub>HPO<sub>4</sub>, 1.8 mM KH<sub>2</sub>PO<sub>4</sub>, 137 mM NaCl, 2.7 mM KCl, pH 7.4) and stored at –80 °C. The identity and purity of the purified proteins were confirmed by MS spectrometry and SDS-PAGE ([Figure S5](#)).

### 6.7.2 Cloning of HL2-m1 Constructs

A DNA sequence encoding for the HL2-m1 sequence fused to an N-terminal FLAG tag and a C-terminal GyrA intein from *Mycobacterium xenopi*[\(28\)](#) (MDYKDDDDK-(GS)<sub>2</sub>-TLDD(*stop*) EEMDGCSD-T-(GyrA)) was assembled by PCR. The resulting gene was cloned into the *Bam*H I/*Xho* I cassette of the expression vector pET22b (Novagen), resulting in the fusion of polyhistidine (H<sub>6</sub>) tag to the C-terminus of the intein. Using a similar procedure, a "self-cleaving" variant[\(22a\)](#) of the FLAG-HL2-m1-GyrA construct

was prepared by removing the Thr residue preceding the intein sequence (= "intein-1" position), thereby leaving an Asp residue at the intein-1 position. In a similar manner, a stable and a self-cleaving variant of the FLAG tag-fused L2-derived peptide (= MDYKDDDDK-(GS)<sub>2</sub>-TLDDMEEMDGLSD-(T)-(GyrA))) were prepared. The sequences of the recombinant vectors were confirmed by DNA sequencing.

### 6.7.3 Library Construction and Screening

The single-site site-saturation libraries were constructed via overlap extension PCR using pET22\_FLAG-(HL2-m1)-D-GyrA as the template and the appropriate mutagenizing primers (NNK codon at target position; forward primers #12–16, #23, #24, reverse primer #8; [Table S1](#)). The PCR product was cloned into the *Bam*H I/ *Xho* I cassette of pET22\_FLAG-(HL2-m1)-D-GyrA. The recombinant plasmids were transformed in DH5α cells and selected on LB plates containing ampicillin (100 µg/mL). The recombination libraries were prepared in a similar manner but using primers with partially randomized codons (codons: KGB, WAM, KGG, WYG, AYG; forward primers #17–22; reverse primer #08) to encode for the desired subset of amino acids at each target position. The resulting plasmid libraries were pooled and transformed into cells containing a pEVOL\_O2beY-RS vector([22a](#)) encoding for the orthogonal O2beY-RS/tRNA<sub>CUA</sub> pair. Recombinant cells were selected on LB plates containing ampicillin (100 µg/mL) and chloramphenicol (34 µg/mL), and individual colonies from these plates were used to inoculate 1.0 mL of LB media containing the two antibiotics in 96-deep well plates. After overnight growth at 37 °C, 50 µL from each well was used to inoculate a replica plate containing 1 mL of M9 medium containing ampicillin (100 µg/mL) and

chloramphenicol (34 µg/mL). Cells were grown to an OD<sub>600</sub> of 0.6 in a plate shaker at 37 °C and then induced with arabinose (0.06% m/v) and O2beY (2 mM). After 1 h, cells were induced with IPTG (1 mM) and grown at 27 °C for 18–20 h. For library screening, the 96-well-plate cell cultures were pelleted by centrifugation and then washed once with PBS buffer. Cell pellets were then resuspended in lysis buffer (50 mM potassium phosphate, 150 mM NaCl, 10 mM MgCl<sub>2</sub>, 0.8 µg/mL DNase, 0.8 mg/mL lysozyme, pH 7.5) and incubated for 1 h and 15 min at 37 °C. After centrifugation, 200 µL of the clarified lysate was used for measuring Shh-binding activity using the immunoassay described further below (see [Hedgehog Binding Assay](#)). Positive hits were identified upon comparison with the reference macrocycle and then validated through rescreening in triplicate using the same overall procedure and assay. The validated hits were then deconvoluted via DNA sequencing of the plasmids extracted from the master plate.

#### 6.7.4 Recombinant Synthesis and Purification of Macrocyclic Peptides

Plasmids for the expression of stable GyrA intein fusions of the macrocyclic peptides were prepared by substituting the Asp residue at the "intein-1" position with Thr in the corresponding pET22-based plasmids via site-directed mutagenesis. The plasmids were cotransformed along with the pEVOL\_O2beY-RS plasmid into *E. coli* BL21(DE3) cells. The recombinant cells were grown in LB media containing ampicillin (100 µg/mL) and chloramphenicol (34 µg/mL) overnight at 37 °C. Overnight cultures were then used to inoculate 1.0 L of M9 media ampicillin (100 µg/mL) and chloramphenicol (34 µg/mL). After growth at 37 °C to an OD<sub>600</sub> of 0.6, the cells were induced with arabinose (0.06% m/v) and O2beY (2 mM). After 1 h, cells were induced with IPTG (1 mM) and grown at

27 °C for 18–20 h. The GyrA-fused peptides were purified by Ni-NTA chromatography (Invitrogen), and the eluted proteins were buffer exchanged with potassium phosphate buffer (10 mM potassium phosphate, 150 mM NaCl, pH 7.5). Cleavage of the intein was carried out using a 200  $\mu$ M solution of purified proteins in potassium phosphate buffer containing 20 mM TCEP (tris(2-carboxyethyl)phosphine) and 10 mM thiophenol at pH 8.5. The reaction mixtures were incubated overnight at room temperature with gentle shaking and then dialyzed against water. The cleaved peptide was purified using solid-phase extraction with a step gradient of acetonitrile in water (+ 0.1% TFA). The peptides generally eluted between 10% and 25% acetonitrile. After lyophilization, the peptides were further purified by reverse-phase HPLC using a Grace C18 column (monomeric; 120 Å; 250  $\times$  10 mm) and a 5%  $\rightarrow$  95% gradient of acetonitrile in water (+ 0.1% TFA (trifluoroacetic acid)). The peptide identity was confirmed by MALDI-TOF MS ([Figure S7](#)), and the concentration was determined by HPLC (OD<sub>220</sub>) using a calibration curve generated with a reference peptide of identical length. Typical yields for the recombinantly produced cyclic peptides obtained using this procedure were between 0.5 and 1.5 mg/L culture.

#### 6.7.5 Molecular Modeling

The peptide variants were mapped onto the backbone scaffold derived from an HHIP L2 loop structure in the Shh-bound state (PDB code: [3HO5](#); [\(25a\)](#) residues 375–387 chain B). On the basis of visual examination of the L2 structure, the backbone psi dihedrals of residue 383 were perturbed by up to 37.5°, and RosettaMatch([35](#)) was used to select backbone conformations that accommodate the O2beY-Cys thioether cross-link between

positions 5 and 11. Geometric constraints to model the preferred geometry of the O2beY-Cys thioether cross-link were derived from the Cambridge Structural Database. The resulting conformations were optimized using the Rosetta FastRelax protocol.<sup>(26)</sup> Total and per-residue energies of the macrocycle residues were used for scoring, and visual examination of models was used to identify favorable interactions. During the energy minimization, geometric constraints were placed on both the metal-chelating Asp10 residue and the O2beY/Cys cross-link. Atom coordinate constraints were placed for backbone atoms of Shh residues outside of the L2 binding cleft to maintain them in their crystallographic conformations. All Rosetta files required to perform simulations are provided as [Supporting Information](#).

#### 6.7.6 Synthesis of O2beY and Dipeptide Building Block

O2beY was synthesized as described previously.<sup>(36)</sup> Detailed procedures for the synthesis of **7** are provided as [Supporting Information](#).

#### 6.7.7 Solid-Phase Peptide Synthesis

The macrocyclic peptides were manually synthesized via standard solid-phase Fmoc chemistry using MBHA (4-methylbenzhydrylamine) rink amide resin (loading: 0.25 mmol/g) in a polypropylene reaction vessel. Standard Fmoc-protected amino acids were used as building blocks, with the exception of Asp10, for which N-Fmoc-Asp(OEpe)CO<sub>2</sub>H was used to avoid aspartimide formation. Loading of the first amino acid and subsequent elongation steps were carried out using 5 equiv of Fmoc-protected amino acid preactivated with COMU ((1-Cyano-2-ethoxy-2-

oxoethylidenaminoxy)dimethylamino-morpholino-carbenium hexafluorophosphate) (4.95 equiv) and diisopropylethylamine (DIPEA) (10 equiv) in dimethylformamide (DMF) for 1 h at room temperature. The Fmoc protecting group was removed with 30% piperidine in DMF ( $2 \times 10$  min). To introduce the dipeptide building block, compound **7** (75 mg 0.11 mmol) was preactivated with COMU/DIPEA in DMF and added to the resin for 3 h at room temperature. Prior to the cyclization reaction, deprotection of the Alloc/allyl groups was carried out using  $\text{Pd}(\text{PPh}_3)_4$  (1 equiv)/ $\text{PhSiH}_3$  (20 equiv) in dry dichloromethane ( $2 \times 45$  min). Peptide cyclization was carried out at millimolar pseudodilution using a mixture of PyBOP (benzotriazol-1-yl-oxytripyrrolidinophosphonium hexafluorophosphate) (2 equiv), HoBt (hydroxybenzotriazole) (2 equiv), and DIPEA (4 equiv) in DMF, for two cycles of 12 h. After addition of the last amino acid in the sequence, the resin-bound peptide was acetylated by two treatments with a mixture of acetic anhydride (0.5 M), DIPEA (0.015 M), and HOBt (0.125 M) in DMF for 10 min. The peptides were cleaved from the solid support using a solution of TFA/ $\text{H}_2\text{O}$ /triisopropylsilane (95:2.5:2.5 v/v/v) for 3 h at room temperature. After removal of the resin by filtration, the crude peptide was precipitated with cold MTBE (methyl-tert-butyl-ether), redissolved in 1:1 water/acetonitrile solution, and lyophilized. The crude peptide was purified by reverse-phase HPLC using an Agilent 1200 system equipped with a Grace C18 column ( $10\ \mu\text{m}$ ;  $90\ \text{\AA}$ ;  $250 \times 10\ \text{mm}$ ) at a flow rate of  $2.5\ \text{mL/min}$  and a linear gradient starting from 20% to 80% acetonitrile in water (+ 0.1% TFA) over 25 min. The purity and identity of all peptide were confirmed by analytical HPLC and LC-MS ([Figures S8 and S9](#)). The overall yield of the macrocyclic peptides obtained using this procedure was around 15%.

### 6.7.8 Hedgehog Binding Assay

Shh-binding activity/affinity of the linear and cyclic peptides was measured using the immunoassay outlined in [Figure 4](#). For these experiments, GST-Shh was immobilized on microtiter plates by incubating 100  $\mu$ L of a 4  $\mu$ M GST-Shh solution in PBS buffer overnight at 4  $^{\circ}$ C, followed by washing ( $3 \times 150$   $\mu$ L of PBS with 0.5% Tween-20) and blocking with 0.5% bovine serum albumin in PBS for 1.5 h at room temperature. After washing, each well was incubated with 200  $\mu$ L of cell lysate for 1 h at room temperature (for library screening). For the  $K_D$  determination experiments ([Figure 3](#)), each well was incubated under the same conditions with 100  $\mu$ L of purified FLAG-fused peptide at varying peptide concentrations. The FLAG-tagged peptides were prepared by recombinant means as described above (see [Recombinant Synthesis and Purification of Macrocyclic Peptides](#)). After washing, each well was incubated with 100  $\mu$ L of 1:2500 dilution of HRP-conjugate mouse anti-FLAG polyclonal antibody (Sigma-Aldrich) for 1 h at room temperature. After washing, 100  $\mu$ L of 2.2 mM *o*-phenylenediamine dihydrochloride, 4.2 mM urea hydrogen peroxide, 100 mM dibasic sodium phosphate, and 50 mM sodium citrate, pH 5.0, were added to each well, followed by measurement of the absorbance at 450 nm after 10–20 min using a Tecan Infinite 1000 plate reader. Equilibrium dissociation constants ( $K_D$ ) were determined by fitting the dose–response curves ([Figure 3](#)) to a 1:1 binding isotherm equation via nonlinear regression using SigmaPlot.  $K_D$  values for HL2-m5 binding to Ihh and Dhh were determined in a similar manner using GST-Ihh- and GST-Dhh-coated plates, respectively. The peptide relative binding affinity for the three analogs of hedgehog ([Figure 7](#)) was determined using the same assay and peptide solutions at a fixed concentration of 0.5–1  $\mu$ M. In this case,

binding responses were subtracted against the blank (no peptide sample) and normalized to the highest value measured across the three Hh analogs. Mean values and standard deviations were calculated from experiments performed at least in triplicate.

#### 6.7.9 Competition Assay

A PBS solution (100  $\mu$ L) containing 10  $\mu$ M robotnikinin and 400 nM FLAG-HL2-m5 was added to GST-Shh-, GST-Ihh-, and GST-Dhh-coated wells in a microtiter plate. The plates were then treated and developed as described above. The relative affinity of robotnikinin for the three Hh analogs was expressed as follows:  $(1 - \% \text{ inhibition})_{\text{GST-Hh}} / (1 - \% \text{ inhibition})_{\text{GST-Shh}}$ , where % inhibition in the presence of GST-Shh was 34%. Mean values and standard deviations were calculated from experiments performed in triplicate. The same assay was applied to determine the  $\text{IC}_{50}$  value for the inhibition of FLAG-HL2-m5 binding to immobilized GST-Shh induced by the synthetic peptide HL2-m5 ([Figure S10](#)).

#### 6.7.10 Circular Dichroism Analyses

CD analyses were performed using solutions of the purified, synthetic peptides at a concentration of 0.4  $\mu$ M in 20 mM potassium phosphate buffer (pH 7) in the absence and in the presence of trifluoroethanol at 50% or 75% (v/v). CD spectra were recorded at 26  $^{\circ}$ C at a scan rate of 50 nm/min with a bandwidth of 1 nm and an averaging time of 10 s per measurement using a JASCO J-1100 CD spectrophotometer. The raw signal ( $\theta_d$ , mDeg) was background subtracted against buffer and converted to molar residue

ellipticity ( $\theta_{\text{MRE}}$ ) using  $\theta_{\text{MRE}} = \theta_d / (c l n_R)$ , where  $c$  is the peptide concentration (M),  $l$  is the path length (1 mm), and  $n_R$  is the number of residues in the peptide.

#### 6.7.11 Serum Stability Assay

The serum stability assay was carried out by dissolving the peptide at a final concentration of 25  $\mu\text{M}$  in 300  $\mu\text{L}$  of 50% human male serum (Sigma-Aldrich) in 20 mM potassium phosphate buffer (pH 7.0). Prior to the assay, the serum was clarified by centrifugation at 14 000 rpm for 15 min and preactivated at 37 °C for 10 min. Each peptide was incubated at 37 °C, and aliquots (45  $\mu\text{L}$ ) were removed over the course of 26 h and quenched with 45  $\mu\text{L}$  of 20% trichloroacetic acid solution, followed by incubation at 4 °C for 15 min and centrifugation at 14 000 rpm for 5 min. The supernatants were analyzed by analytical RP-HPLC (Grace Vision HT C18 HL column; 21.2  $\times$  250 mm; 5  $\mu\text{m}$ ) using a gradient from 10% to 75% of acetonitrile (0.1% TFA) in water (0.1% TFA) at a flow rate of 1 mL/min. The residual peptide was quantified based on the corresponding peak area (210 nm) as relative to the sample at time zero. Mean values and standard deviations were calculated from experiments performed in triplicate.

#### 6.7.12 Gli-Reporter Assay

NIH3T3 cells (AATC CRL-1658) were passaged twice and then plated in 24-well culture dishes at  $5 \times 10^5$  cells/well in Dulbecco's modified Eagle's medium (DMEM) containing 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin. After 24 h, the cells were transfected (TransIT-2020) with a mixture of a firefly luciferase reporter construct under the control of a Gli1 inducible promoter and a Renilla luciferase reporter construct under

a constitutive promoter (40:1) (Cignal GLI reporter luciferase kit, Qiagen). Cells were allowed to reach confluency, at which point the media was changed to Opti-MEM containing 1% FBS and added with 4 nM Shh-N (R&D Systems, Minneapolis, MN, USA) in sterile PBS buffer. Synthetic HL2-pep and HL2-m5 were added at the same time at varying concentrations (0.01–30  $\mu$ M), and control cells were prepared by adding vehicle only (1% DMSO). Purmorphamine-treated cells were prepared by adding 5  $\mu$ M purmorphamine to wells containing 10  $\mu$ M HL2-m5. After growth for 24 h at 37 °C in a humidified chamber, the cells were harvested and analyzed for firefly and Renilla luciferase activity using a Tecan Spark-20 plate reader and a DLR kit (Promega) according to the manufacturer's instructions. Luminescence values were normalized to those of the Shh pathway activated control cells. Mean values and standard deviations were calculated from experiments performed at least in duplicate.

#### 6.7.13 Gene Transcription Analyses

NIH3T3 cells were passaged twice and plated at a density of 1:3 in DMEM containing 10% FBS and 1% penicillin/streptomycin in six-well cell culture dishes. Cells were allowed to reach confluency, at which point the media was changed to Opti-MEM containing 1% FBS and added with Shh-N (4 nM). At the same time, the cells were incubated with HL2-m5 (25  $\mu$ M) or vehicle only (1% DMSO). After growth for 24 h at 37 °C in a humidified chamber, the cells were harvested, and total mRNA was collected using TRIzol reagent (ThermoFisher) according to the manufacturer's instructions. cDNA was generated using 1  $\mu$ g of mRNA using First-Strand RT-PCR with random hexamers (Super-Script First-Strand RT-PCR, ThermoFisher). The relative amounts of

*Gli1*, *Gli2*, and *Ptch1* mRNA transcripts were determined by real-time PCR (Bio-Rad CFX thermocycler) using the primers listed in [Table S1](#) and SYBR green TAQ reagent (Bio-Rad) according to the manufacturer's protocol. The mRNA levels for the biomarker genes were normalized to that of the reference house-keeping gene cyclophilin. Mean values and standard deviations were calculated from measurements performed in quadruplicate.

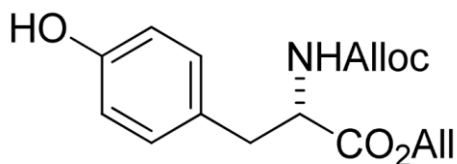
## 6.8 Supporting Information

### 6.8.1 Materials and Methods

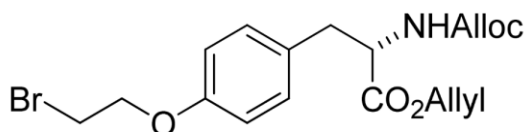
#### 6.8.1.1 General Information

Chemical reagents and solvents were purchased from Sigma–Aldrich, Acros Organics, and Fluka and used without further purification unless stated otherwise. Rink Amide MBHA resin, activating reagents (COMU, PyBop and HOBt), Fmoc-protected amino acids, L-Tyrosine allyl ester (pToluene sulfonate salt) and L-Cystine tert-butyl ester were purchased from Chemimpex. Fmoc-Asp(OEpe)-OH was purchased from Novabiochem. Silica gel chromatography purifications were carried out by using AMD Silica Gel 60 230–4nd00 mesh. <sup>1</sup>H and <sup>13</sup>C NMR spectra were recorded on Bruker Avance spectrometers by using solvent peaks as reference. LC-MS analyses were performed on a Thermo Scientific LTQ Velos ESI/ion-trap mass spectrometer coupled to an Accela U-HPLC system. MALDITOF spectra were acquired on a Bruker Autoflex III MALDI-TOF spectrometer by using a stainless steel MALDI plate and sinapinic acid or alpha-cyano-4-hydroxycinnamic acid (CHCA) as matrix.

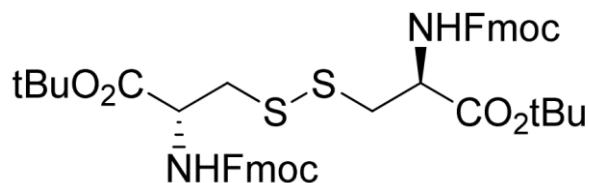
## 6.8.1.2 Synthetic Procedures



*Synthesis of N-Alloc-L-Tyrosine allyl ester (2).* L-Tyrosine allyl ester (pToluene sulfonate salt) (1) (1.7 g, 4.32 mmol) was dissolved in 15 mL of water. Sodium carbonate was added to the solution (1,361 g, 12.96 mmol, 3 equiv), then allyl chloroformate (6.48 mmol, 0.68 mL, 1.5 equiv) was added dropwise to the reaction at 0°C. The reaction was stirred for 15 hours, after which it was quenched by addition of 1 M HCl (15 mL) and extracted with ethyl acetate (2 x 40 mL). The combined organic layers were washed with water (70 mL) and dried over Na<sub>2</sub>SO<sub>4</sub>. After removal of the solvent by rotary evaporation, the crude product was purified on a silica gel column using hexanes/ethyl acetate from 9:1 to 8:2 as eluent to yield 2 as a colorless oil (0.92 g, 70%). <sup>1</sup>H-NMR (400MHz, MeOD) δ 6.99-6.97 (d, 2H, J = 8.4 Hz), 6.67-6.65 (d, 2H, J = 8.4 Hz), 5.88-5.79 (m, 2H, J = 6.4 Hz), S3 5.26-5.09 (m, 4H, J = 9.4 Hz), 4.54-4.53 (d, 2H, J= 5.6 Hz), 4.45-4.44 (d, 2H, J= 4.8 Hz), 4.34 (t, 1H, J= 6.0 Hz), 3.01-2.79 (m, 2H, J = 8.8 Hz). <sup>13</sup>C-NMR (100 MHz, MeOD) δ 171.7, 171.4, 156.7, 155.8, 132.7, 131.7, 129.7, 127.3, 117.1, 116.0, 114.7, 65.2, 64.9, 59.9, 55.7, 36.3, 19.4, 12.9 MS-ESI: Calc. Mass for C<sub>16</sub>H<sub>19</sub>NO<sub>5</sub>: 305.3 Da. Obs. Mass for [MH]<sup>-</sup> : 304.3 Da.

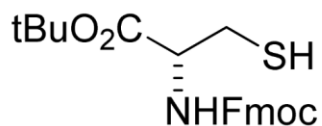


*Synthesis of N-Alloc O-(2-bromoethyl)-L-Tyrosine allyl ester (3).* N-Alloc-L-tyrosine allyl ester 2 (0.92 g, 3.02 mmol) was dissolved in 15 mL dry DMF under argon flow. K<sub>2</sub>CO<sub>3</sub> (1.25 g, 9.06 mmol, 3 equiv) was added to the reaction and stirred vigorously for 10 minutes. Then 1,2-dibromoethane (0.8 mL, 9.06 mmol, 3 equiv) was added to the suspension dropwise. The reaction was stirred overnight and then quenched with HCl 1M (15 mL). The crude product was extracted using ethyl acetate (2 x 40 mL). The combined organic layers were washed with brine and dried over Na<sub>2</sub>SO<sub>4</sub>. After removal of the solvent by rotary evaporation, the crude product was purified on silica gel column using hexanes/ethyl acetate from 9:1 to 7:3 to yield 3 as a colorless oil (0.43 mg, 35%) and recovered starting material (0.55 g, 60%). <sup>1</sup>H-NMR (400MHz, MeOD) δ 7.02-7.00 (d, 2H, J= 8.4 Hz) 6.80-6.78 (d, 2H, J= 8.8 Hz), 5.89-5.78 (m, 2H, J= 6.0 Hz), 5.29-5.15 (m, 4H, J= 10.0 Hz), 4.57-4.56 (d, 2H, J= 5.6 Hz), 4.52-4.50 (d, 2H, J= 5.2 Hz), 4.57-4.51 (m, 1H), 4.23-4.20 (t, 2H, J= 6.4 Hz), 3.59-3.56 (t, 2H, J= 6.4 Hz), 3.08-2.97 (m, 2H, J= 6.4 Hz) <sup>13</sup>C-NMR (100MHz, MeOD) δ 171.1, 157.1, 155.3, 132.4, 131.3, 130.3, 128.4, 118.9, 117.6, 114.7, 67.7, 65.9, 65.6, 54.7, 37.2, 28.9. ESI-MS: Calc. Mass for C<sub>18</sub>H<sub>22</sub>BrNO<sub>5</sub>: 412.28 Da. Obs. Mass for [M+Na]<sup>+</sup>: 434.3 Da. S4



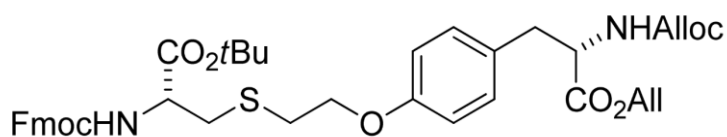
*Synthesis of (2R,2'R)-di-tert-butyl 3,3'-disulfanedibis(2-(((9H-fluoren-9-yl)methoxy)carbonyl)amino)propanoate (5).* L-Cystine tert-butyl ester 4 (2 mmol, 704

mg) was suspended in 10 mL of THF and N-methyl morpholine (4 mmol, 0.520 mL, 2 equiv) was added to the suspension. The solution was chilled to 0°C in an ice bath and then 9-fluorenylmethyl-N-succinimidyl carbonate (Fmoc-OSu) (2 mmol, 675 mg) was added slowly portion-wise. The reaction was stirred for 18 hours allowing to return at room temperature. The solvent was removed under reduced pressure and the crude product was dissolved in 25 mL of ethyl acetate. The organic layer was washed with 20 mL of HCl 0.1 M and then with 20 ml of brine. The organic layer was dried over Na<sub>2</sub>SO<sub>4</sub> filtered and evaporated. The crude product was purified by silica gel column using hexanes/diethyl ether (7:3) to yield 5 as a white solid (1.2 g, 75%). <sup>1</sup>H-NMR (400MHz, CDCl<sub>3</sub>) 7.76-7.74 (d, 4H, J= 7.6 Hz), 7.62-7.60 (d, 4H, J= 7.2 Hz), 7.41-7.37 (t, 2H, J= 7.2 Hz), 7.32-7.29 (t, 2H, J= 7.2 Hz), 4.48-4.46 (m, 2H), 4.39-4.38 (t, 4H, J= 7.2 Hz), 4.25-4.23 (t, 2H, J= 7.2 Hz), 3.24-3.15 (m, 4H), 1.47 (s, 9H). <sup>13</sup>C-NMR (100MHz, CDCl<sub>3</sub>) 169.4, 155.5, 143.6, 127.5, 126.9, 125.0, 125.0, 119.8, 82.9, 79.8, 67.1, 60.2, 46.9, 28.6, 28.2, 27.8, 27.3. ESIMS. Calc. Mass for C<sub>44</sub>H<sub>48</sub>N<sub>2</sub>O<sub>8</sub>S<sub>2</sub>: 796.29 Da Obs. Mass: 819.4 [M+Na].



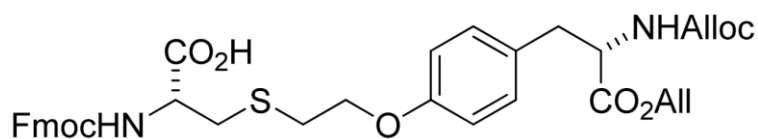
*Synthesis of N-Fmoc-L-Cysteine t-butyl ester (6).* 1.2 g of N,N'-Fmoc-Cystine t-butyl ester (5) (1.72 mmol) was dissolved in 20 mL of THF. Triphenylphosphine (0.9 g, 3.44 mmol, 2 equiv) was added to the solution and the reaction mixture was stirred for 2 hours at room temperature. Water (2 mL) was then added and the reaction mixture was stirred

for 10 hours. The solvent was removed by rotary evaporation and the residue was taken up in EtOAc, washed with 10% citric acid and brine, dried over sodium sulfate and concentrated. The crude product was purified on silica gel column using hexanes/ethyl acetate from 95:5 to 8:2 ratio to yield 6 as a colorless oil (0.4 g, 60%). <sup>1</sup>H-NMR (400MHz, S5 CDCl<sub>3</sub>) 7.78-7.76 (d, 2H, J= 7.6 Hz), 7.62-7.60 (d, 2H, J= 7.2 Da), 7.42-7.39 (t, 2H, J= 7.6 Da), 7.34-7.30 (t, 2H, J= 7.6 Hz), 4.54 (m, 1H), 4.43-4.39 (t, 2H, J= 7.2 Hz), 4.25-4.21 (t, 1H, J= 6.8 Hz), 3.00-2.98 (m, 2H, J= 7.6 Hz), 1.43 (s, 9H). <sup>13</sup>C-NMR (100MHz, CDCl<sub>3</sub>) 171.0, 159.9, 141.1, 127.5, 127.4, 126.9, 125.0, 124.6, 119.8, 82.5, 68.2, 67.1, 60.2, 46.9, 27.9, 27.7, 20.8. MS-ESI- : Calc. Mass for C<sub>22</sub>H<sub>25</sub>NO<sub>4</sub>S: 399.51 Da Obs. Mass: 422.3 [M+Na].



(*R*)-*tert*-butyl 2-((((9*H*-fluoren-9-yl)methoxy)carbonyl)amino)-3-((2-(4-((*S*)-2-(((allyloxy)carbonyl)amino)-3-oxo-3-(prop-1-en-1-yloxy)propyl)phenoxy)ethyl)thio)propanoate (6*b*). N-alloc-O-(2-bromoethyl)-L-Tyrosine allyl ester 3 (0.43 mg, 1.04 mmol) and N-Fmoc-L-cysteine *t*-butyl ester 6 (0.41 mg, 1.04 mmol) were dissolved in 5 mL of dry ethyl acetate. Tetrabutylammonium bromide (1.29 g, 4.0 mmol) was dissolved in 5 mL of nitrogen-sparged NaHCO<sub>3</sub> solution (0.5 M), which was added to the reaction mixture dropwise under argon. The reaction was stirred vigorously for 16 hours, then diluted with ethyl acetate. The organic layer was washed with water and brine, dried over Na<sub>2</sub>SO<sub>4</sub> and concentrated. The crude product was purified on silica gel column using hexanes/ethyl acetate from 9:1 to 7:3 to yield 7 as a colorless oil (0.29 g, 40%). <sup>1</sup>H-NMR

(400MHz, CDCl<sub>3</sub>)  $\delta$  7.73-7.71 (d, 2H, J= 7.2 Hz), 7.57-7.55 (d, 2H, J= 7.6 Hz), 7.37-7.34 (t, 2H, J= 7.6 Hz), 7.28-7.24 (t, 2H, J= 8.0 Hz), 6.99-6.96 (d, 2H, J= 8.4 Hz), 6.78-6.75 (d, 2H, J= 8.4 Hz), 5.82 (m, 2H, J= 6.4 Hz), 5.29-5.15 (m, 4H, J= 10.8 Hz), 4.57-4.55 (d, 2H, J= 5.6 Hz), 4.52-4.50 (d, 2H, J= 5.2 Hz), 4.35-4.34 (t, 2H, J= 3.6 Hz), 4.18 (t, 1H, J= 6.8 Hz), 4.07 (t, 2H, J= 7.2 Hz), 3.10-2.99 (m, 4H, J= 5.2 Hz), 2.90-2.87 (t, 2H, J= 6.0 Hz), 1.45 (s, 9H). <sup>13</sup>C-NMR (100MHz, CDCl<sub>3</sub>)  $\delta$  171.1, 169.5, 157.39, 155.5, 155.3, 143.6, 141.1, 131.2, 130.2, 127.9, 127.5, 126.9, 124.9, 119.8, 118.9, 117.7, 114.5, 82.8, 67.6, 66.9, 65.8, 65.6, 60.2, 54., 54.2, 46.9, 37.1, 35.0, 31.7, 27.8, 14.0. S6



*Synthesis of (R)-2-((((9H-fluoren-9-yl)methoxy)carbonyl)amino)-3-((2-(4-((S)-2-(((allyloxy)carbonyl)amino)-3-oxo-3-(prop-1-en-1-yloxy)propyl)phenoxy)ethyl)thio)propanoic acid (7).* To a solution of 6b (0.29 g, 0.4 mmol) in CH<sub>2</sub>Cl<sub>2</sub> (6 mL) was added 4 mL of trifluoroacetic acid (TFA) at 0°C. The reaction was stirred at 0°C for 2 hours. The product was concentrated in vacuo, then washed extensively with diethyl ether. The final product was yielded as a white crystalline powder. <sup>1</sup>H-NMR (400MHz, MeOD)  $\delta$  7.77- 7.75 (d, 2H, J= 7.2 Hz), 7.65-7.63 (d, 2H, J= 6.8 Hz), 7.37-7.34 m (t, 2H, J= 7.6 Hz), 7.29- 7.25 (t, 2H, J= 7.6 Hz), 7.08-7.057 (d, 2H, J= 8.4 Hz), 6.82-6.79 (d, 2H, J= 8.8 Hz), 5.85 (m, 2H), 5.29-5.11 (m, 4H), 4.57-4.30 (m, 8H, J= 5.6 Hz), 4.19 (t, 1H), 4.11-4.08 (t, 2H, J= 6.0 Hz), 2.99-2.84 (m, 6H). <sup>13</sup>C-NMR (100MHz, MeOD)  $\delta$  173.5, 171.5, 171.4, 169.8, 157.6, 157.4, 156.1, 156.0, 155.8, 143.7, 141.3, 132.4, 131.7, 131.3, 130.4,

128.1, 128.0, 127.7, 127.1, 125.1, 124.8, 120.0, 119.2, 119.1, 118.3, 114.8, 114.7, 83.2, 67.9, 67.3, 66.2, 55.9, 55.0, 47.1, 37.4, 35.9, 31.9.

### 6.8.2 Supporting Figures and Tables

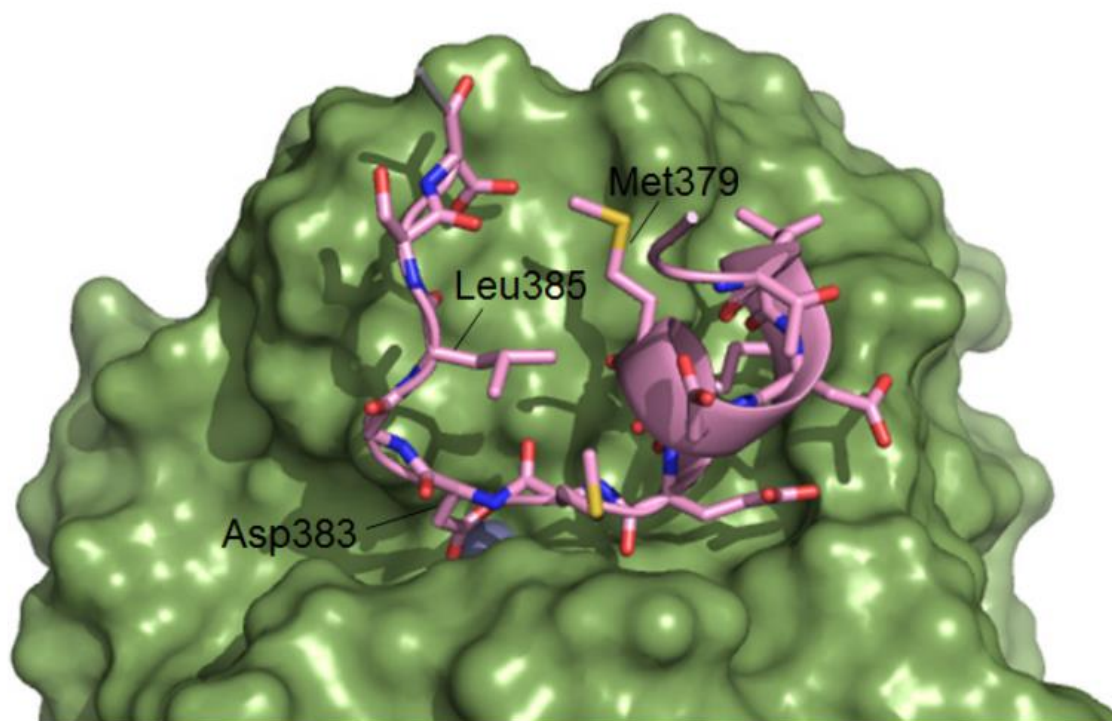
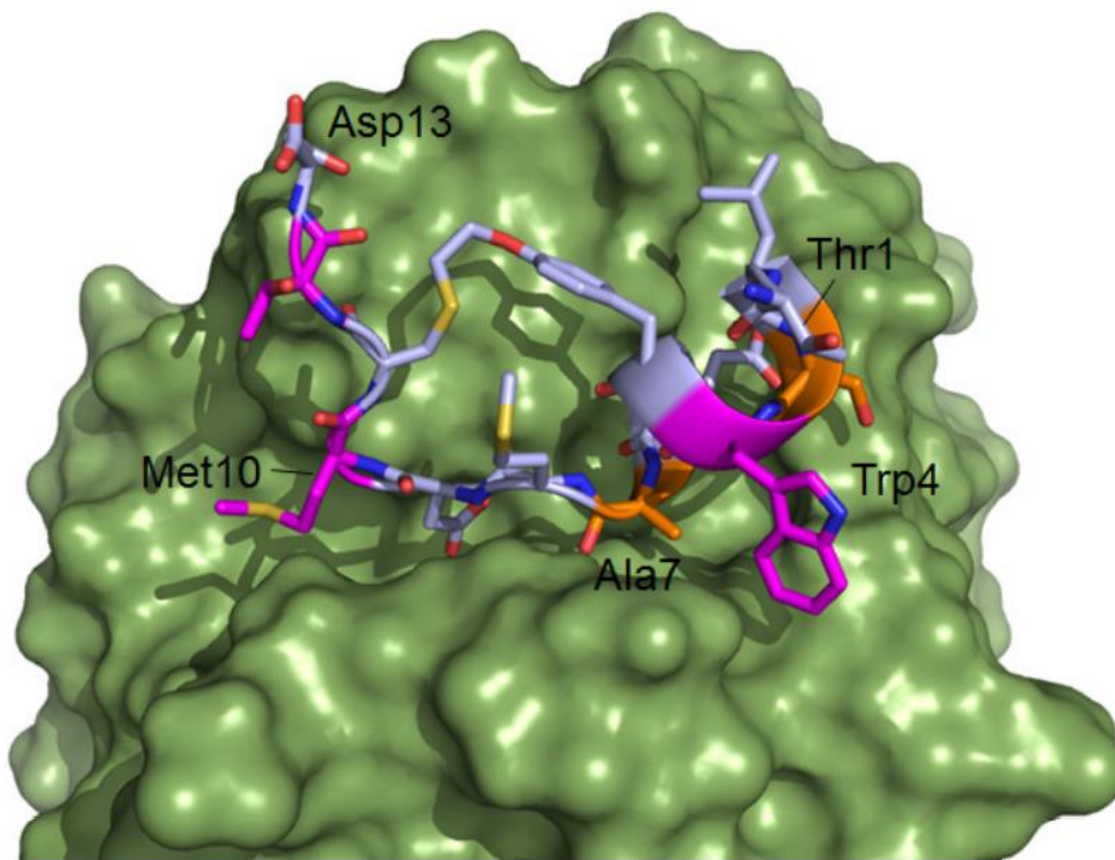


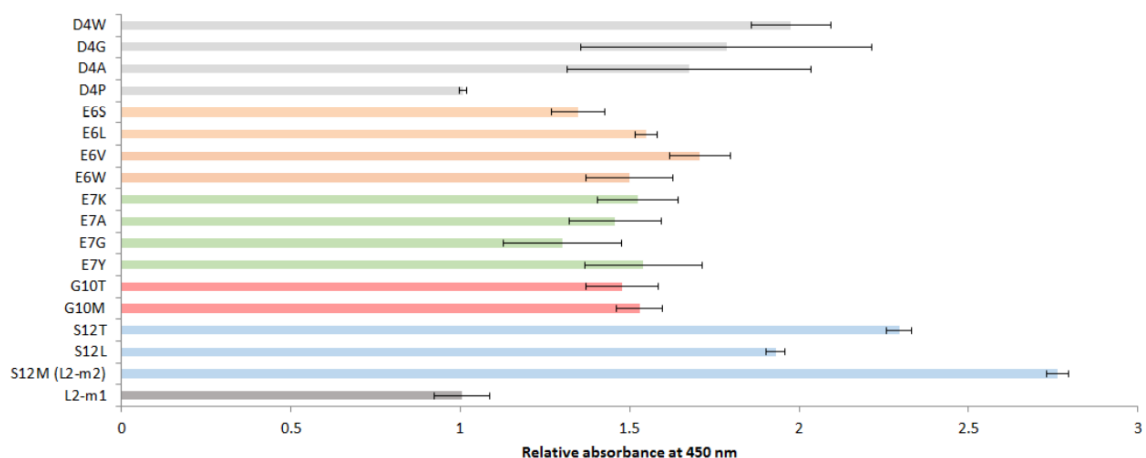
Figure 65-6 Figure S1. Close-up view of HHIP L2 loop interaction with Shh (pdb 3HO5).

The Shh protein is shown as a surface model (green), whereas the L2 loop region of HHIP is shown as a stick model (pink). The remainder of the HHIP protein as shown in Figure 2 is omitted for clarity. The residues selected for the installation of the thioether bridge along with the zinc ion binding aspartate residue are labeled. The zinc ion is shown as sphere model (blue).

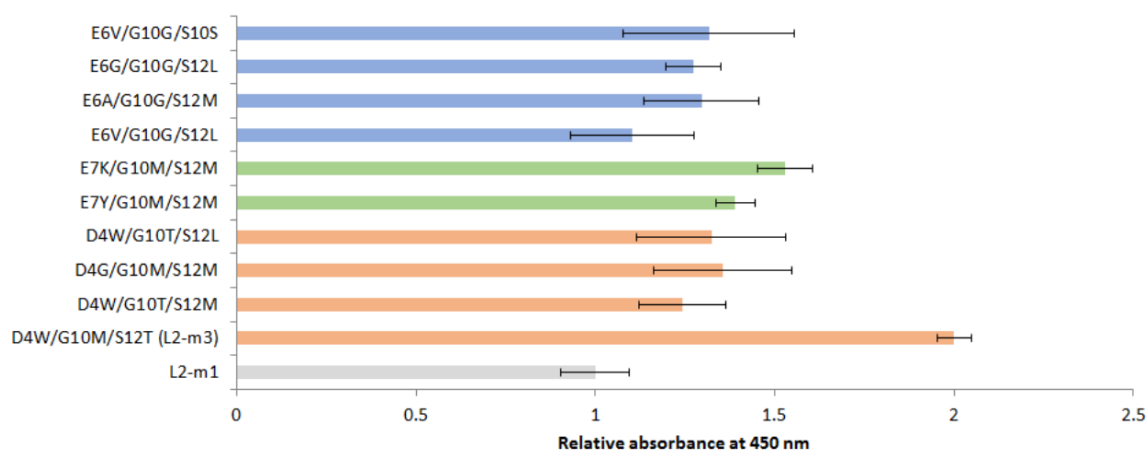


**Figure 66-6.** Figure S2. Model of evolved macrocyclic peptide HL2-m5 in complex with Shh.

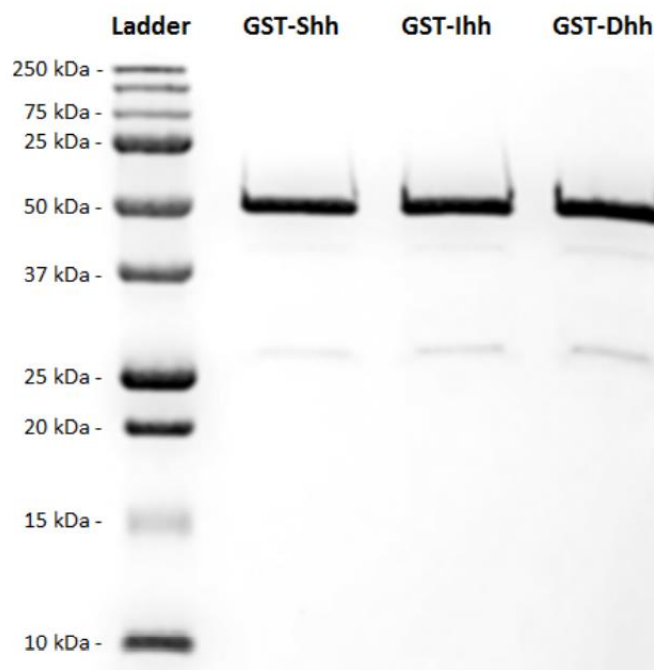
Shh protein is shown as a surface model (green), whereas the macrocyclic peptide is shown as a stick model (blue). The mutated residues with respect to HL2-m1 are color coded as shown in Figure 3. The N-terminal and C-terminal residues along with Trp4, Ala7, and Met10 are labeled.



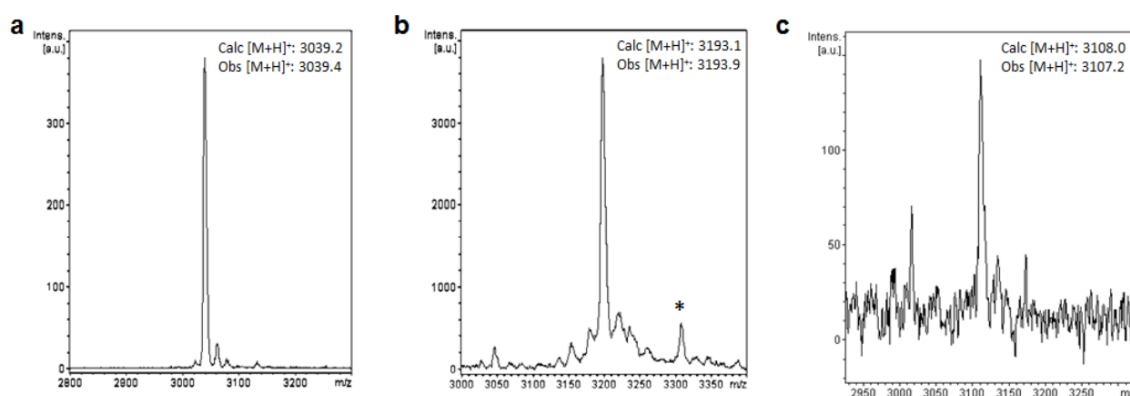
**Figure 67-6.** Figure S3. Relative Shh binding activity for representative hits from the single-site sitesaturation libraries. Absorbance values (X axis) are normalized to that of HL2-m1. Indicated mutations (Y axis) are relative to the HL2-m1 sequence. The mean values and error bars were derived from rescreening of the hits identified during the library screening in triplicate.



**Figure 68-6.** Figure S4. Relative Shh binding activity for representative hits from the multi-site recombinant libraries. Absorbance values (X axis) are normalized to that of HL2-m1. Indicated mutations (Y axis) are relative to the HL2-m1 sequence. The mean values and error bars were derived from rescreening of the hits identified during the library screening in triplicate.

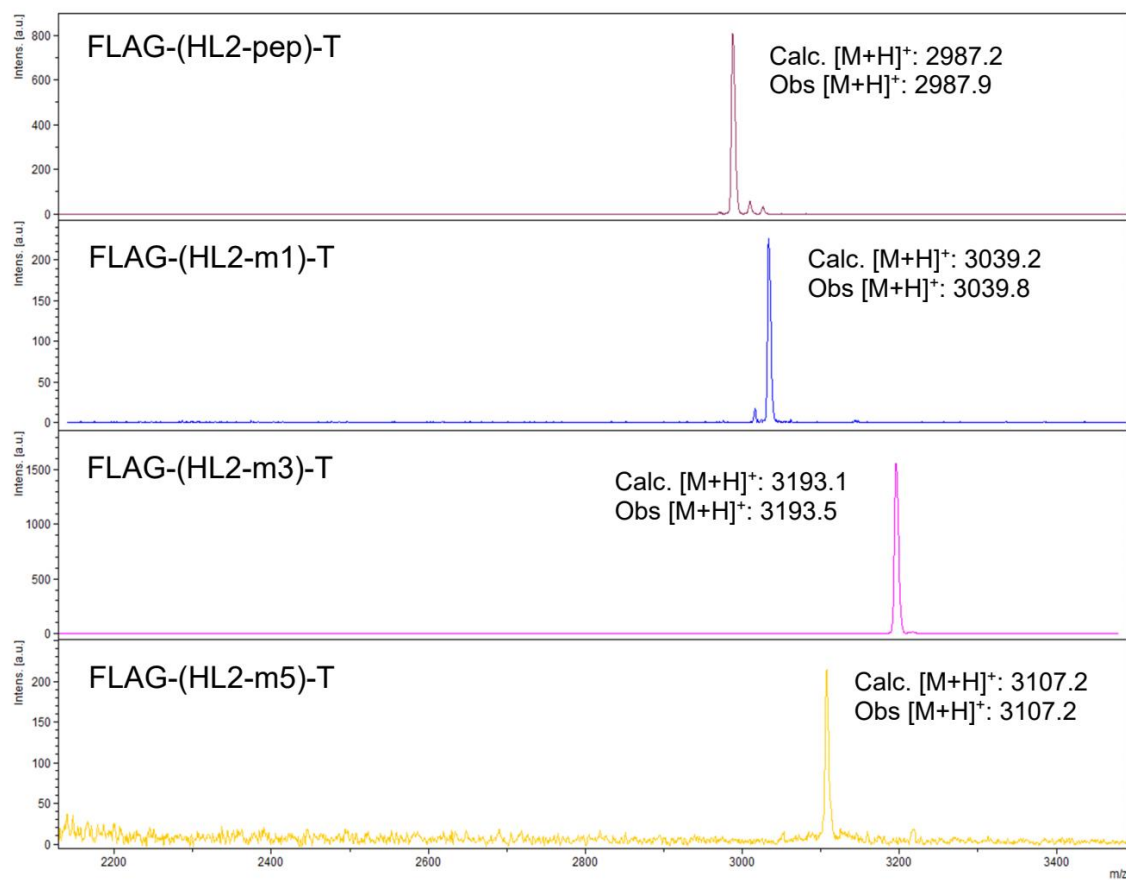


**Figure 69-6.** Figure S5. SDS-PAGE gel of recombinantly expressed GST-Shh, GST-Ihh, and GST-Dhh after purification by Ni-affinity chromatography.



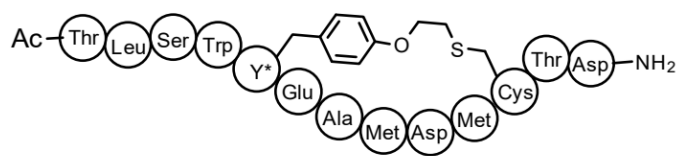
**Figure 70-6.** Figure S6. Thiol-induced intein cleavage reactions.

MALDI-TOF MS spectra corresponding to the GyrA cleavage reactions for FLAG-(HL2-m1)-GyrA (a), FLAG-(HL2-m3)-GyrA (b), and FLAG-(HL2-m5)-GyrA (c) after incubation with thiophenol. Calculated and observed  $m/z$  values corresponding to the proton adducts of the macrocycles are indicated. The species labeled with the star (\*) corresponds to thiophenol thioester. The absence of acyclic or hydrolysis byproducts indicates that the constructs have undergone quantitative cyclization upon expression in *E. coli* cells.

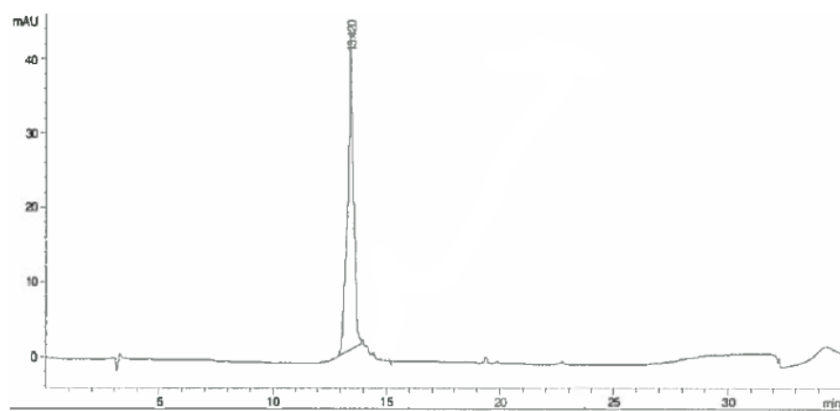


**Figure 71-6.** Figure S7. MALDI-TOF MS spectra corresponding to purified FLAG-tagged linear and cyclic L2 mimics obtained via recombinant expression.

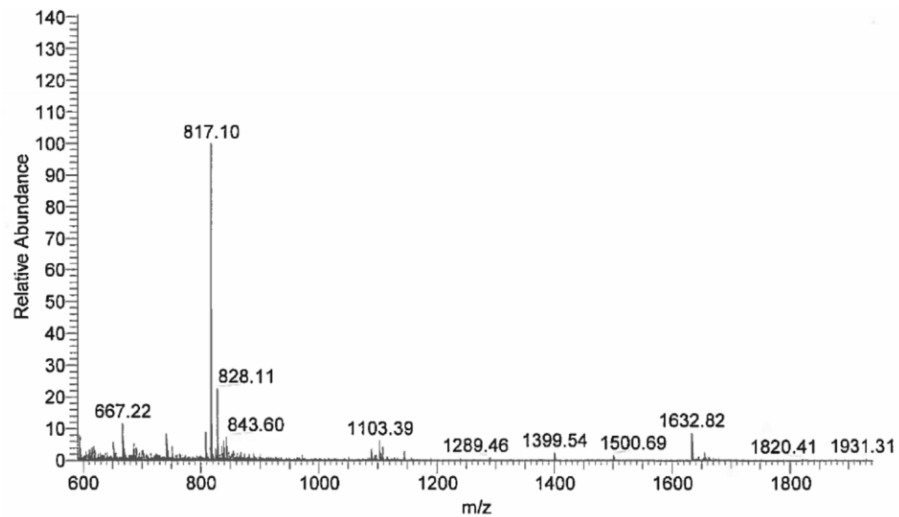
Calculated and observed m/z values corresponding to the proton adduct of the macrocycles are indicated.

**HL2-m5**

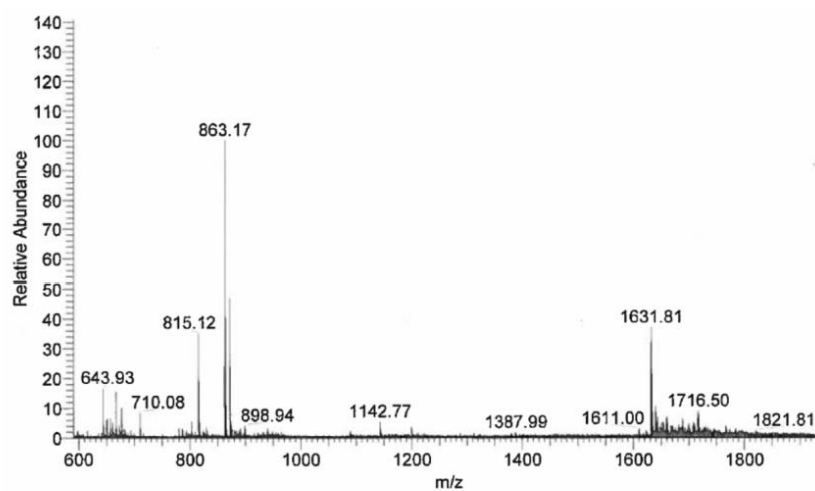
A)



B)

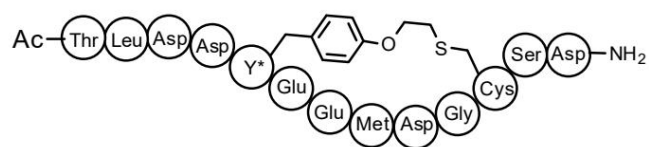


C)

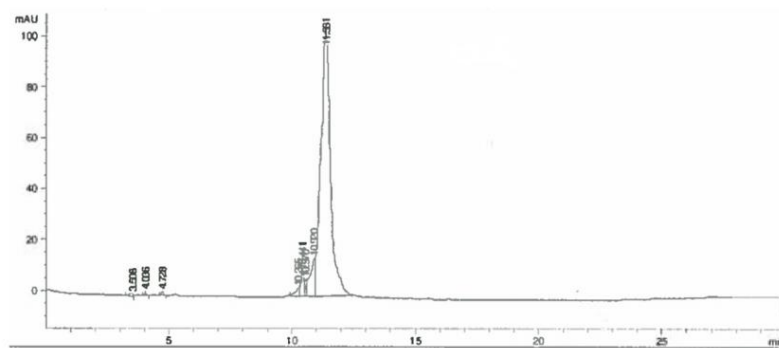


**Figure 72-6.** Figure S8. Analytical HPLC chromatogram (A) and ESI-MS spectra in positive (B) and negative mode (C) corresponding to synthetic HL2-m5. Y\* = alkylated O2beY.

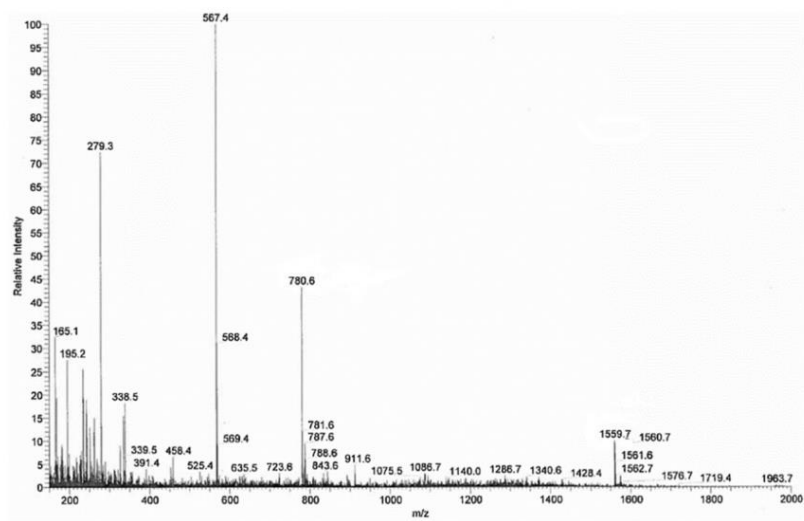
See Table S2 for further details.

**HL2-m1**

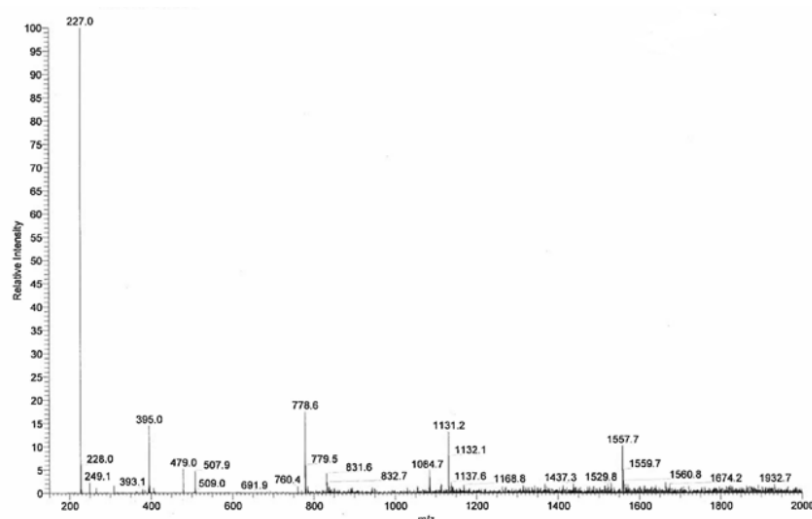
A)



B)

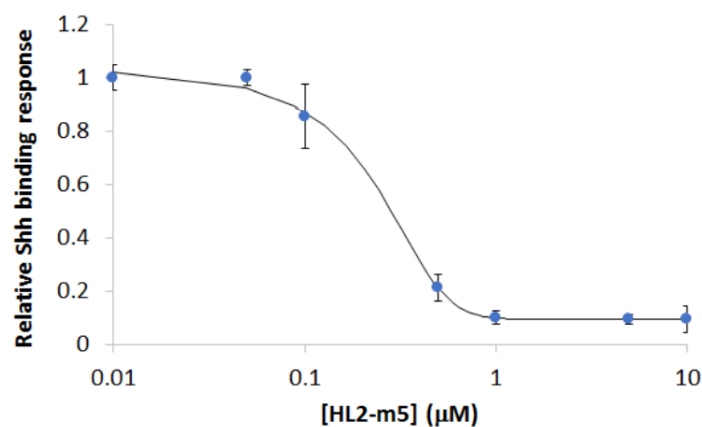


C)



**Figure 73-6.** Figure S9. Analytical HPLC chromatogram (A) and ESI-MS spectra in positive (B) and negative mode (C) corresponding to synthetic HL2-m1. Y\* = alkylated O2beY.

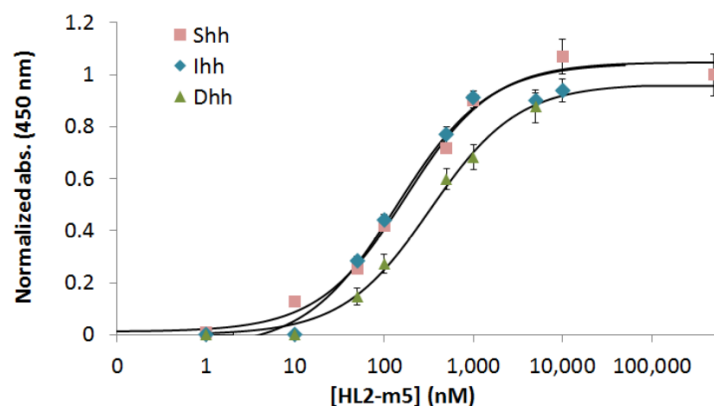
See Table S2 for further details.



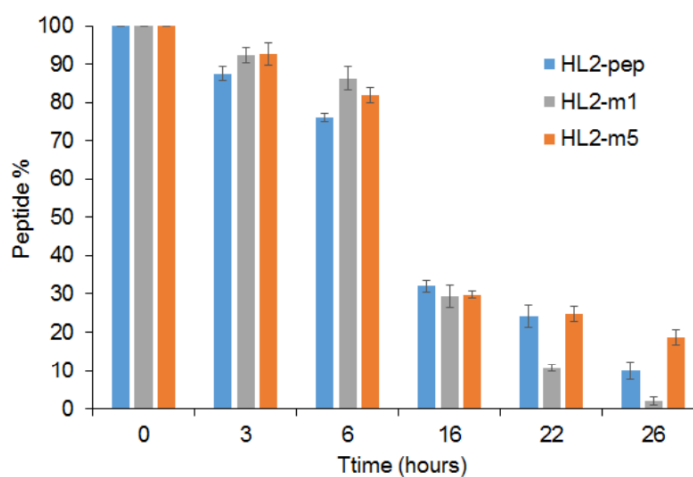
**Figure 74-6.** Figure S10. Inhibition curve corresponding to HL2-m5 induced inhibition of FLAG-HL2- m5 binding to plate-immobilized GST-Shh.

The data were fitted to a four-parameter equation, from which a  $IC_{50}$  of  $280 \pm 50$  nM was calculated. The mean values and standard deviations were obtained from experiments performed in triplicate. The similarity between the  $IC_{50}$  value

determined in this assay and the  $K_D$  value measured for FLAG-HL2-m5 in the direct Shh binding assay (Figure 3) indicates the FLAG tag does not significantly affect the Shh binding affinity of the cyclic peptide.



**Figure 75-6.** Figure S11. Dose-response curves for direct binding of FLAG-HL2-m5 to plate-immobilized GST-Shh, GST-Ihh, or GST-Dhh as determined using the colorimetric assay with HRP-conjugated anti-FLAG antibody.



**Figure 76-6.** Figure S12. Proteolytic stability of linear and cyclic L2 mimics.

The graph indicates the residual amount of HL2-pep, HL2-m1, and HL2-m5 peptides related after incubation in human blood serum (37°C) at different time points as determined by analytical HPLC. Values are normalized to peak areas corresponding to the same peptide in buffer only. Under identical assay conditions, an unrelated linear peptide (p5315–29 peptide in Smith et al., Chemical Commun. 2014, 50, 5027) exhibited a half-time ( $t_{1/2}$ ) < 1 hour, indicating the HHIP L2-derived sequence is inherently resistant to proteolytic degradation.

Table 11-6. Table S1. Oligonucleotide Sequences

Primer	Sequence (5' to 3')
01_Shh_Forward	CTGCGCCATGGGTGGACCGGGCAGGGGGT
02_Shh-Reverse	GAAGACTCGAGTCAGCCTCCCGATTTGGCCG
03_Dhh_Forward	ACTATACCATGGGTGGGCCGGGCCG
04_Dhh_Reverse	ACTATACTCGAGTCAGCCCGCCCGAC
05_Ihh_Forward	ACTATACCATGGGTGGGCCGGGTCTGGGTGGT
06_Ihh_Reverse	ACTATACTCGAGTCAGCCCGTCTTGGCTGCGG
07_L2(T)_Forward	TAGAGGATCCACCCTGGACGATATGGAAGAGATGGACGGCCTGAGTGA TACCTGCATCACGG
08_GyrA reverse	CAAAAAACCCCTCAAGACCCGTTTAGAGGCCCAAGGGGTTATGCTA
09_L2(D)_Forward	TAGAGGATCCACCCTGGACGATATGGAAGAGATGGACGGCCTGAGTGA TGATTGCATCACGG
10_L2-m1(T)_Forward	TAGAGGATCCACCCTGGACGATTAGGAAGAGATGGACGGCTGCAGTGA TACCTGCATCACGG
11_L2-m1(D)_Forward	TAGAGGATCCACCCTGGACGATTAGGAAGAGATGGACGGCTGCAGTGA TGATTGCATCACGG
12_D4(NNK)_Forward	TAGAGGATCCACCCTGGACNNKTAGGAAGAGATGGACGGCTGCAGTGA TTGCATCACGGG
13_E6(NNK)_Forward	TAGAGGATCCACCCTGGACGATTAGNNKGAGATGGACGGCTGCAGTGA TTGCATCACGGG
14_E7(NNK)_Forward	TAGAGGATCCACCCTGGACGATTAGGAANNKATGGACGGCTGCAGTGA TTGCATCACGGG
15_G10(NNK)_Forward	TAGAGGATCCACCCTGGACGATTAGGAAGAGATGGACNNKTCAGTGA TTGCATCACGGG
16_S12(NNK)_Forward	TAGAGGATCCACCCTGGACGATTAGGAAGAGATGGACGGCTGCNNKGA TTGCATCACGGG
17_Recombination1_F1	TAGAGGATCCACCCTGGACGATTAGKBGGAGATGGACAYGTGCWYGGA TTGCATCACGGG
18_Recombination1_F2	TAGAGGATCCACCCTGGACGATTAGKBGGAGATGGACGGCTGCWYGGA TTGCATCACGGG
19_Recombination2_F1	TAGAGGATCCACCCTGGACGATTAGGAAWAMATGGACAYGTGCWYGG ATTGCATCACGGG
20_Recombination2_F2	TAGAGGATCCACCCTGGACGATTAGGAAWAMATGGACGGCTGCWYGG ATTGCATCACGGG
21_Recombination3_F1	TAGAGGATCCACCCTGGACKGGTAGGAAGAGATGGACAYGTGCWYGG ATTGCATCACGGG
22_Recombination3_F2	TAGAGGATCCACCCTGGACGATTAGGAAGAGATGGACGGCTGCWYGGA TTGCATCACGGG
23_L2-m3_D3(NNK)_Forward	TAGAGGATCCACCCTGNNKTGGTAGGAAGAGATGGACATGTGCACCGA TACCTGCATCAC
24_L2-m3_E7(NNK)_Forward	TAGAGGATCCACCCTGGATTGGTAGGAANNKATGGACATGTGCACCGA TACCTGCATCAC
25_L2-m3(T)	TAGAGGATCCACCCTGGATTGGTAGGAAGAGATGGACATGTGCACCGA TACCTGCATCAC
26_L2-m5(T)	TAGAGGATCCACCCTGTCCTGGTAGGAAGCCATGGACATGTGCACCGAT ACCTGCATCAC
27_Cyclophilin F*	TATAAGGGTTCCTCCTTCACAGAA
28_Cyclophilin R*	GGACCTGTATGCTTTAGGATGAAGT
29_Gli1F*	AAGGAATTCGTGTGCCATTGGG
30_Gli1R*	ACATGTAAGGCTTCTACCCGT

31_Gli2F*	TCCAGTCAATGGTTCTGTCC
32_Gli2R*	TGGCTCAGCATCGTCACTTC
33_Ptch1F*	CATAGCTGCCCAGTTCAAGT
34_Ptch1R*	GGTCGTAAAGTAGGTGCTGG

(\*) Denotes real-time PCR primer.

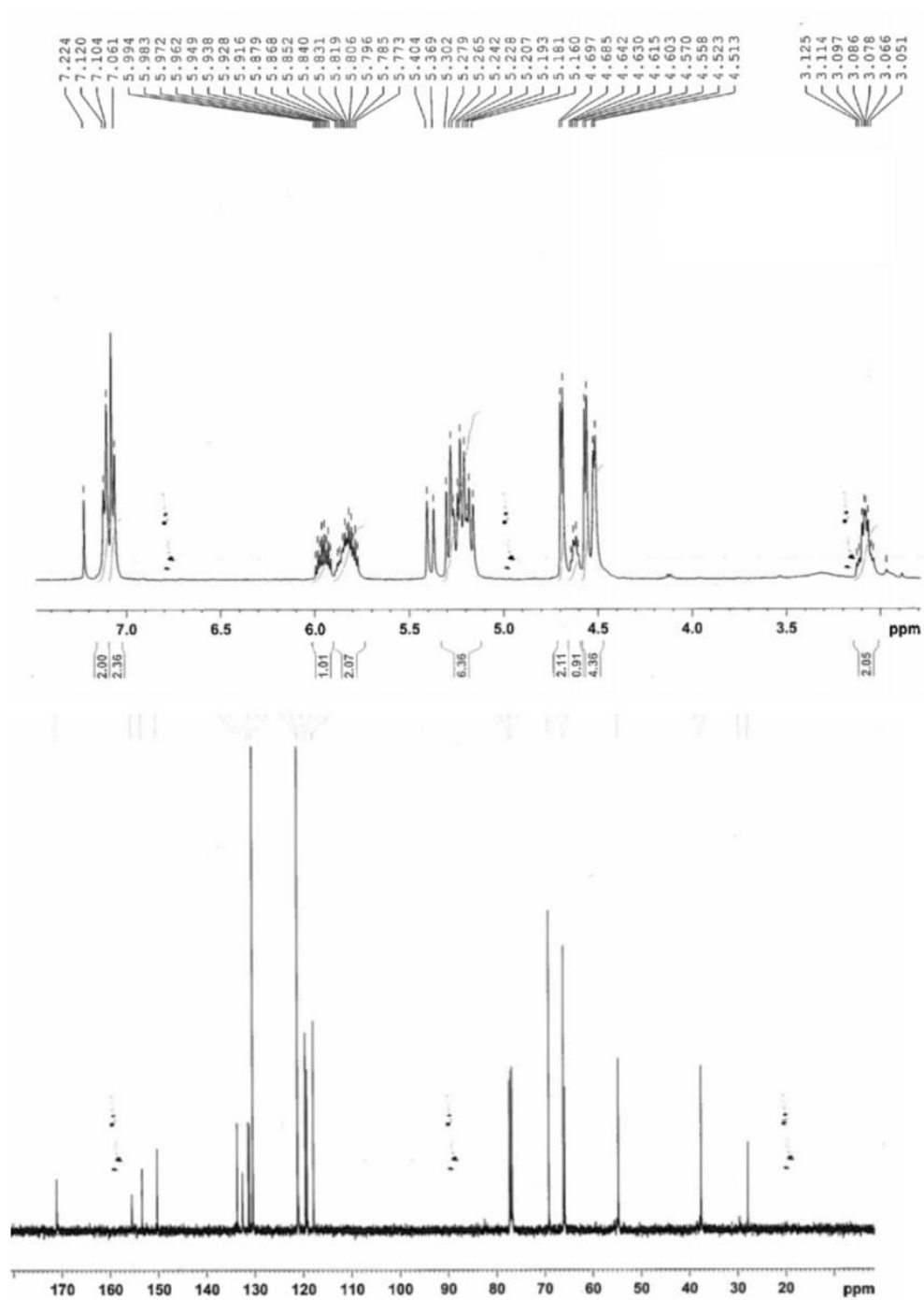
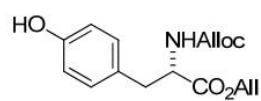
Table 12-6. S2 MS data and retention times for linear and cyclic L2 mimics.

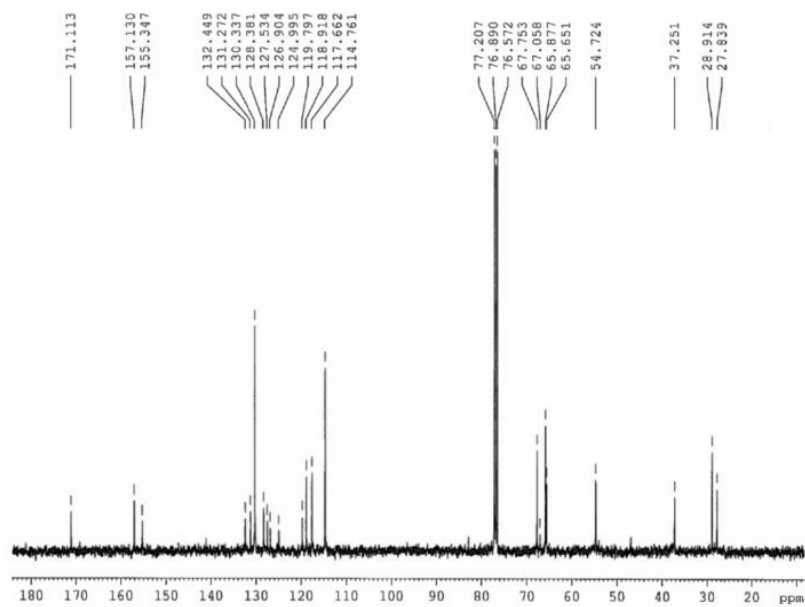
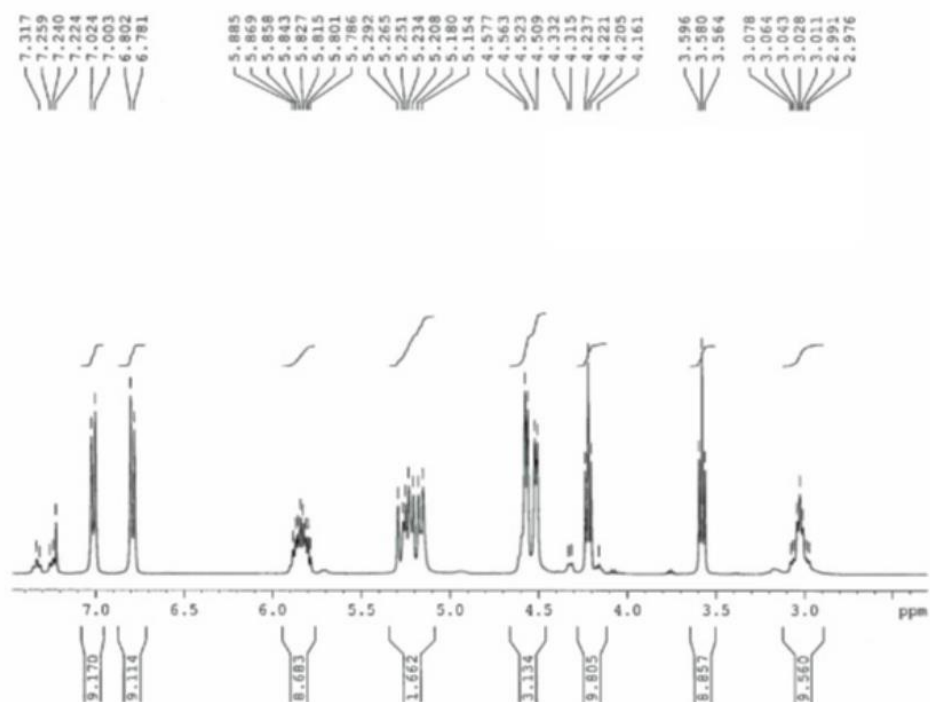
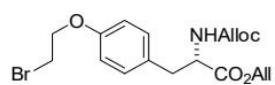
HPLC analyses were performed using an Agilent 1200 series HPLC system equipped with a Grace Vision HT C18 HL column (21.2 x 250 mm; 5 $\mu$ ) and multidiode array detector. Method: linear gradient of H<sub>2</sub>O (0.1% TFA)/CH<sub>3</sub>CN (0.1% TFA) from 20 to 75% of CH<sub>3</sub>CN (0.1% TFA) in 17 min at a flow rate of 1 mL/min.

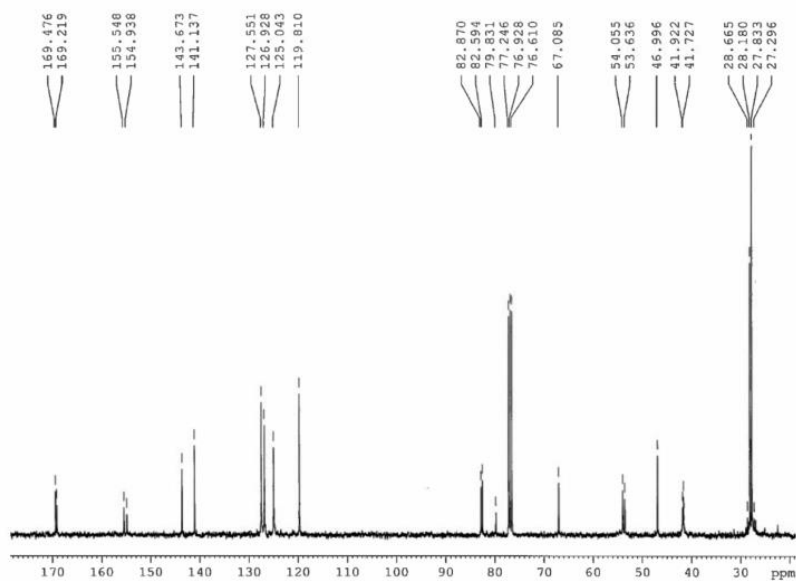
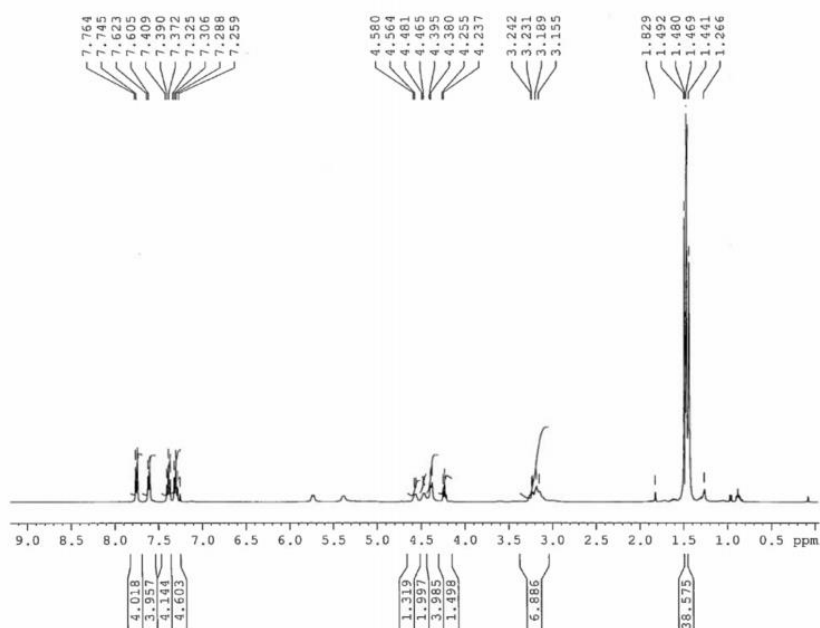
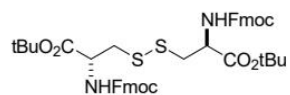
Peptide	Calc. Mass	Observed Mass [M+H] <sup>+</sup>	Observed Mass [M+Na] <sup>+</sup>	Observed Mass [M-H] <sup>-</sup>	Retention Time
HL2-pep	1511.59 Da	1512.44 Da	Not obs.	n/a	9.9 min
HL2-m1	1559.59 Da	1559.70 Da	Not obs.	1557.70 Da	11.3 min
HL2-m5	1632.84 Da	1633.82 Da	1656.40 Da	1631.81 Da	13.42 min

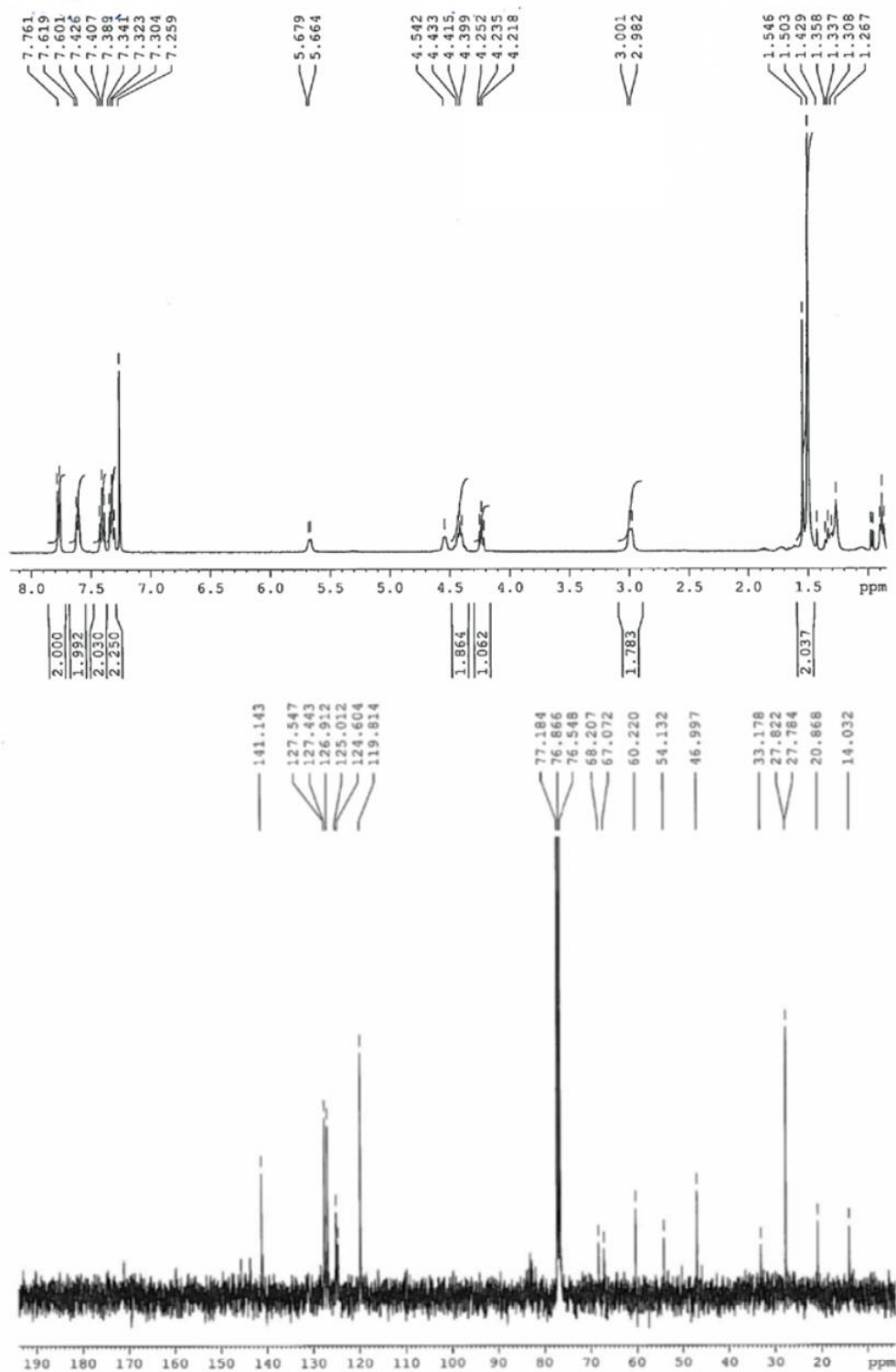
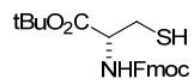
## 6.8.2.1 NMR Spectra

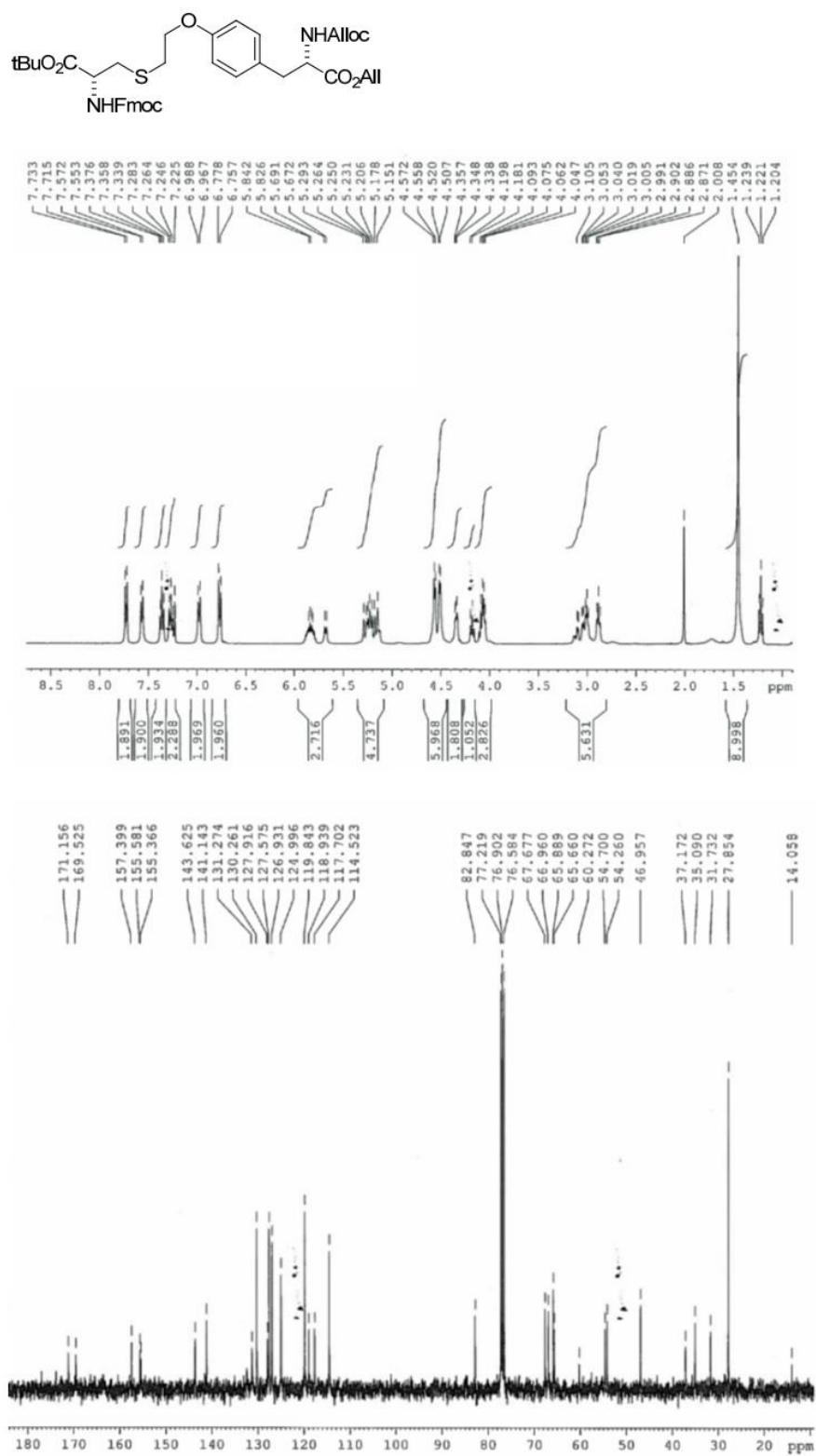
## NMR spectra

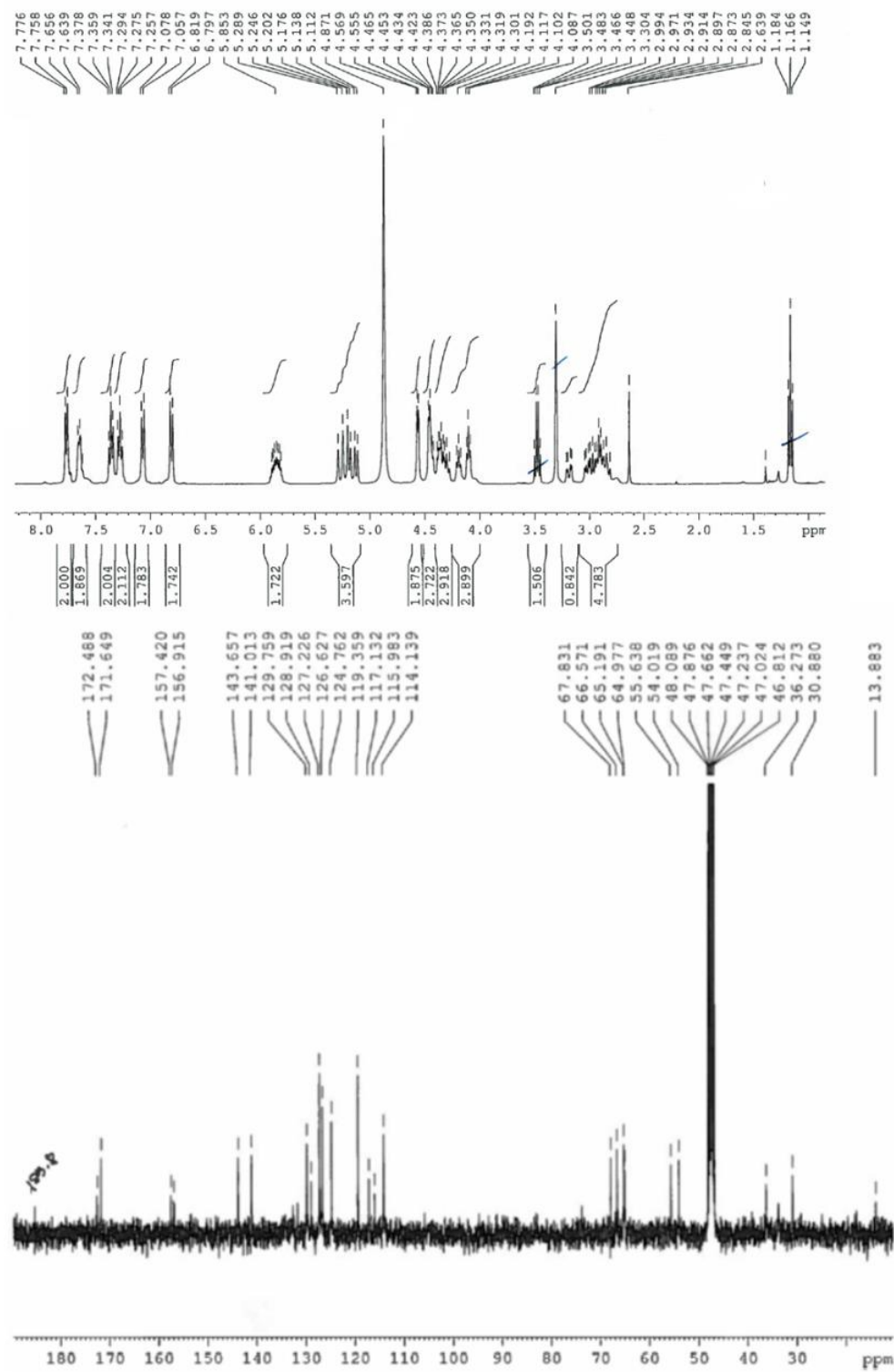
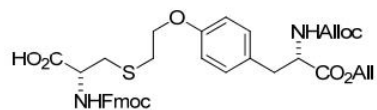












### 6.8.2.2 Rosetta Files

#### Rosetta files

```
CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: C1 C2 O1
  TEMPLATE:: ATOM_MAP: 1 residue3: XLK

  TEMPLATE:: ATOM_MAP: 2 atom_type: S ,
  TEMPLATE:: ATOM_MAP: 2 residue3: CYZ

  CONSTRAINT:: distanceAB: 1.82 0.05 50.0 1 0
  CONSTRAINT:: angle_A: 109.4 5.0 15.0 360.0 0
  CONSTRAINT:: angle_B: 102.0 5.0 20.0 360.0 0
  CONSTRAINT:: torsion_A: 180.0 30.0 10.0 120.0 3
  CONSTRAINT:: torsion_B: 180.0 30.0 10.0 120.0 3
  CONSTRAINT:: torsion_AB: 180.0 30.0 10.0 120.0 3
  ALGORITHM_INFO:: match
    IGNORE_UPSTREAM_PROTON_CHI
    CHI_STRATEGY:: CHI 1 EX_TWO_HALF_STEP_STDDEVS
  ALGORITHM_INFO::END
CST::END
```

```
CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: C2 C1 S1
  TEMPLATE:: ATOM_MAP: 1 residue3: XLK

  TEMPLATE:: ATOM_MAP: 2 atom_type: OH CZ CE1
  TEMPLATE:: ATOM_MAP: 2 residue3: TYZ

  CONSTRAINT:: distanceAB: 1.42 0.03 50.0 1 0
  CONSTRAINT:: angle_A: 109.4 3.0 15.0 360.0 0
  CONSTRAINT:: angle_B: 117.0 3.0 20.0 360.0 0
  CONSTRAINT:: torsion_A: 180.0 30.0 10.0 120.0 3
  CONSTRAINT:: torsion_B: 0. 10.0 10.0 180.0 1
  CONSTRAINT:: torsion_AB: 180.0 30.0 10.0 120.0 3
  ALGORITHM_INFO:: match
    IGNORE_UPSTREAM_PROTON_CHI
  ALGORITHM_INFO::END
CST::END
```

Rosetta relax XML:

```
<ROSETTASCRIPTS>
  <SCOREFXNS>
    <myscore weights=talaris2013_cst >
      <Reweight scoretype=atom_pair_constraint weight=1.0 />
      <Reweight scoretype=angle_constraint weight=1.0 />
      <Reweight scoretype=dihedral_constraint weight=1.0 />
      <Reweight scoretype=coordinate_constraint weight=1.0 />
    </myscore>
  </SCOREFXNS>

  <RESIDUE_SELECTORS>

    <Index name=peptide resnums=156-167 />
    <Not name=not_peptide selector=peptide />

  </RESIDUE_SELECTORS>

  <TASKOPERATIONS>
    <InitializeFromCommandline name=init/>
    <IncludeCurrent name=keep_curr/>

    <OperateOnResidueSubset name=relaxPeptide selector=peptide >
      <PreventRepackingRLT/>
    </OperateOnResidueSubset>
```

```

    <OperateOnResidueSubset name=relaxRestWithCSTs selector=not_peptide >
      <RestrictToRepackingRLT/>
    </OperateOnResidueSubset>

</TASKOPERATIONS>

<FILTERS>
</FILTERS>

<MOVERS>

  <AddOrRemoveMatchCsts name=enzCST cst_instruction="add_new" cstfile="binding.cst" keep_covalent=1
/>
  <AtomCoordinateCstMover name=floppyPeptide coord_dev=0.2 bounded=true bound_width=0.1
sidechain=true native=false task_operations=relaxPeptide />
  <FastRelax name=fastrelax repeats=8 scorefxn=myscore
task_operations=keep_curr,init,relaxPeptide,relaxRestWithCSTs />
  <LoopOver name=fast5 mover_name=fastrelax iterations=5 drift=true/>

</MOVERS>

<APPLY_TO_POSE>
</APPLY_TO_POSE>

<PROTOCOLS>
  <Add mover=floppyPeptide />
  <Add mover=enzCST />
  <Add mover=fast5 />
</PROTOCOLS>
</ROSETTASCRIPTS>

```

### Binding.cst:

```

CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: C2 C1 S1
  TEMPLATE:: ATOM_MAP: 1 residue3: XLK

  TEMPLATE:: ATOM_MAP: 2 atom_type: OH ,
  TEMPLATE:: ATOM_MAP: 2 residue3: TYZ

  CONSTRAINT:: distanceAB: 1.43 0.05 50.0 1
  CONSTRAINT:: angle_A: 109.4 5.0 15.0 360.0
  CONSTRAINT:: angle_B: 105.0 5.0 15.0 360.0
  CONSTRAINT:: torsion_A: 180.0 30.0 10.0 120.0
  CONSTRAINT:: torsion_B: 0. 10.0 10.0 180.0
  CONSTRAINT:: torsion_AB: 180.0 30.0 10.0 120.0
CST::END

CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: C1 C2 O1
  TEMPLATE:: ATOM_MAP: 1 residue3: XLK

  TEMPLATE:: ATOM_MAP: 2 atom_type: S ,
  TEMPLATE:: ATOM_MAP: 2 residue3: CYZ

  CONSTRAINT:: distanceAB: 1.82 0.05 50.0 1
  CONSTRAINT:: angle_A: 109.4 5.0 10.0 360.0
  CONSTRAINT:: angle_B: 95.0 5.0 15.0 360.0
  CONSTRAINT:: torsion_A: 180.0 30.0 10.0 120.0
  CONSTRAINT:: torsion_B: 180.0 30.0 10.0 120.0
  CONSTRAINT:: torsion_AB: 180.0 30.0 10.0 120.0
CST::END

```

CST::BEGIN  
 TEMPLATE:: ATOM\_MAP: 1 atom\_name: ZN CA1 CA2  
 TEMPLATE:: ATOM\_MAP: 1 residue3: CAZ

TEMPLATE:: ATOM\_MAP: 2 atom\_type: NE2 CD2 CG  
 TEMPLATE:: ATOM\_MAP: 2 residue3: HIS

CONSTRAINT:: distanceAB: 2.069 0.05 50.0 1  
 CONSTRAINT:: angle\_A: 57.5 5.0 15.0 360.0  
 CONSTRAINT:: angle\_B: 122.7 5.0 15.0 360.0  
 CONSTRAINT:: torsion\_A: -55.4 10.0 10.0 360.0  
 CONSTRAINT:: torsion\_B: -173.0 10.0 10.0 360.0  
 CONSTRAINT:: torsion\_AB: 123.7 10.0 10.0 360.0

CST::END

CST::BEGIN  
 TEMPLATE:: ATOM\_MAP: 1 atom\_name: ZN CA1 CA2  
 TEMPLATE:: ATOM\_MAP: 1 residue3: CAZ

TEMPLATE:: ATOM\_MAP: 2 atom\_type: ND1 CG CB  
 TEMPLATE:: ATOM\_MAP: 2 residue3: HIS

CONSTRAINT:: distanceAB: 2.073 0.05 50.0 1  
 CONSTRAINT:: angle\_A: 150.4 5.0 15.0 360.0  
 CONSTRAINT:: angle\_B: 135.8 5.0 15.0 360.0  
 CONSTRAINT:: torsion\_A: -116.7 10.0 10.0 360.0  
 CONSTRAINT:: torsion\_B: 2.4 10.0 10.0 360.0  
 CONSTRAINT:: torsion\_AB: -102.1 10.0 10.0 360.0

CST::END

CST::BEGIN  
 TEMPLATE:: ATOM\_MAP: 1 atom\_name: ZN CA1 CA2  
 TEMPLATE:: ATOM\_MAP: 1 residue3: CAZ

TEMPLATE:: ATOM\_MAP: 2 atom\_type: OD1 CG CB  
 TEMPLATE:: ATOM\_MAP: 2 residue3: ASP

CONSTRAINT:: distanceAB: 1.997 0.05 50.0 1  
 CONSTRAINT:: angle\_A: 56.7 5.0 15.0 360.0  
 CONSTRAINT:: angle\_B: 120.5 5.0 15.0 360.0  
 CONSTRAINT:: torsion\_A: 174.4 10.0 10.0 360.0  
 CONSTRAINT:: torsion\_B: 171.2 10.0 10.0 360.0  
 CONSTRAINT:: torsion\_AB: -21.6 10.0 10.0 360.0

CST::END

CST::BEGIN  
 TEMPLATE:: ATOM\_MAP: 1 atom\_name: ZN CA1 CA2  
 TEMPLATE:: ATOM\_MAP: 1 residue3: CAZ

TEMPLATE:: ATOM\_MAP: 2 atom\_type: OD2 CG CB  
 TEMPLATE:: ATOM\_MAP: 2 residue3: ASP

CONSTRAINT:: distanceAB: 2.168 0.05 50.0 1  
 CONSTRAINT:: angle\_A: 109.0 5.0 15.0 360.0  
 CONSTRAINT:: angle\_B: 111.3 5.0 15.0 360.0  
 CONSTRAINT:: torsion\_A: 38.4 10.0 10.0 360.0  
 CONSTRAINT:: torsion\_B: -178.0 10.0 10.0 360.0  
 CONSTRAINT:: torsion\_AB: 64.5 10.0 10.0 360.0

CST::END

CST::BEGIN  
 TEMPLATE:: ATOM\_MAP: 1 atom\_name: ZN CA1 CA2  
 TEMPLATE:: ATOM\_MAP: 1 residue3: CAZ

TEMPLATE:: ATOM\_MAP: 2 atom\_type: OD1 CG OD2  
 TEMPLATE:: ATOM\_MAP: 2 residue3: ASP

```

CONSTRAINT:: distanceAB: 2.869  0.05 50.0  1
CONSTRAINT::  angle_A: 85.4   5.0 15.0 360.0
CONSTRAINT::  angle_B: 77.4   5.0 15.0 360.0
CONSTRAINT::  torsion_A: 82.1  30.0 10.0 360.0
CONSTRAINT::  torsion_B: -1.8  10.0 10.0 360.0
CONSTRAINT::  torsion_AB: -119.0 30.0 10.0 360.0
CST::END

```

#Calcium csts with OOCs start here

```

CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: CA2 ZN CA1
  TEMPLATE:: ATOM_MAP: 1 residue3: CAZ

```

```

  TEMPLATE:: ATOM_MAP: 2 atom_name: OD1 CG CB
  TEMPLATE:: ATOM_MAP: 2 residue3: ASP

```

```

  CONSTRAINT:: distanceAB: 2.321  0.01 50.0  1
CST::END

```

```

CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: CA2 ZN CA1
  TEMPLATE:: ATOM_MAP: 1 residue3: CAZ

```

```

  TEMPLATE:: ATOM_MAP: 2 atom_name: OD2 CG CB
  TEMPLATE:: ATOM_MAP: 2 residue3: ASP

```

```

  CONSTRAINT:: distanceAB: 2.335  0.01 50.0  1
CST::END

```

```

CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: CA1 CA2 ZN
  TEMPLATE:: ATOM_MAP: 1 residue3: CAZ

```

```

  TEMPLATE:: ATOM_MAP: 2 atom_type: OOC ,
  TEMPLATE:: ATOM_MAP: 2 residue3: ASP

```

```

  CONSTRAINT:: distanceAB: 2.331  0.01 50.0  1
CST::END

```

```

CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: CA1 CA2 ZN
  TEMPLATE:: ATOM_MAP: 1 residue3: CAZ

```

```

  TEMPLATE:: ATOM_MAP: 2 atom_type: OOC ,
  TEMPLATE:: ATOM_MAP: 2 residue3: GLU

```

```

  CONSTRAINT:: distanceAB: 2.335  0.01 50.0  1
CST::END

```

```

CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: CA1 CA2 ZN
  TEMPLATE:: ATOM_MAP: 1 residue3: CAZ

```

```

  TEMPLATE:: ATOM_MAP: 2 atom_type: OOC ,
  TEMPLATE:: ATOM_MAP: 2 residue3: GLU

```

```

  CONSTRAINT:: distanceAB: 2.343  0.01 50.0  1
CST::END

```

```

CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: CA1 CA2 ZN
  TEMPLATE:: ATOM_MAP: 1 residue3: CAZ

```

```

  TEMPLATE:: ATOM_MAP: 2 atom_name: OE2 CD CG

```

```

TEMPLATE:: ATOM_MAP: 2 residue3: GLU

CONSTRAINT:: distanceAB: 2.336  0.01  50.0  1
CST::END

CST::BEGIN
TEMPLATE:: ATOM_MAP: 1 atom_name: CA1 CA2 ZN
TEMPLATE:: ATOM_MAP: 1 residue3: CAZ

TEMPLATE:: ATOM_MAP: 2 atom_name: O CA N
TEMPLATE:: ATOM_MAP: 2 residue3: THR

CONSTRAINT:: distanceAB: 2.355  0.01  50.0  1
CST::END

CST::BEGIN
TEMPLATE:: ATOM_MAP: 1 atom_name: CA2 CA1 ZN
TEMPLATE:: ATOM_MAP: 1 residue3: CAZ

TEMPLATE:: ATOM_MAP: 2 atom_type: OOC ,
TEMPLATE:: ATOM_MAP: 2 residue3: GLU

CONSTRAINT:: distanceAB: 2.327  0.01  50.0  1
CST::END
~

```

### Run command:

```
~/Rosetta/main/source/bin/rosetta_scripts.linuxgccrelease @general.flags -s $1
```

## **Chapter 7: Structures of the peptide-modifying radical SAM enzyme SuiB elucidate the basis of substrate recognition**

### **7.1 Preface**

A version of this chapter has been published in the *Proceedings of the National Academy of Sciences USA* and is formatted in the journal style.

### **7.2 Abstract**

Posttranslational modification of ribosomally synthesized peptides provides an elegant means for the production of biologically active molecules known as RiPPs (ribosomally synthesized and posttranslationally modified peptides). Although the leader sequence of the precursor peptide is often required for turnover, the exact mode of recognition by the modifying enzymes remains unclear for many members of this class of natural products. Here, we have used X-ray crystallography and computational modeling to examine the role of the leader peptide in the biosynthesis of a homolog of streptide, a recently identified peptide natural product with an intramolecular lysine–tryptophan cross-link, which is installed by the radical *S*-adenosylmethionine (SAM) enzyme, StrB. We present crystal structures of SuiB, a close ortholog of StrB, in various forms, including apo SuiB, SAM-bound SuiB, and a complex of SuiB with SAM and its peptide substrate, SuiA. Although the N-terminal domain of SuiB adopts a typical RRE (RiPP recognition element) motif, which has been implicated in precursor peptide recognition, we observe binding of the leader peptide in the catalytic barrel rather than the N-terminal domain. Computational simulations support a mechanism in which the leader peptide guides

posttranslational modification by positioning the cross-linking residues of the precursor peptide within the active site. Together the results shed light onto binding of the precursor peptide and the associated conformational changes needed for the formation of the unique carbon–carbon cross-link in the streptide family of natural products.

### 7.3 Introduction

Peptide natural products have had a profound impact on human health as sources of antibacterial, anticancer, and antifungal therapeutics ([1](#), [2](#)). Broadly speaking, they may be synthesized in a ribosome-dependent or -independent manner. The former category comprises the family of ribosomally-synthesized and posttranslationally modified peptides (RiPPs). Recent advancements in genome sequencing and bioinformatics have led to the rapid discovery of a multitude of RiPPs and their biosynthetic gene clusters ([3](#)). Unlike nonribosomal peptides, which are assembled by large multimodular enzymes ([4](#)), RiPP biosynthetic pathways are comparatively simple and thus attractive targets for bioengineering ([3](#)). Biosynthesis commences with the ribosomal production of a precursor peptide whose core sequence is modified by tailoring enzymes. Proteolytic removal of the N- and/or C-terminal portions of the peptide, which occurs in most studied cases, followed by export of the mature product, completes RiPP biogenesis. RiPPs encompass structurally and chemically diverse subclasses, such as lanthipeptides, cyanobactins, thiopeptides, and sactipeptides ([3](#)). Most recently, a new subclass of RiPPs was identified with the discovery of streptide. It contains an unprecedented lysine–tryptophan carbon–carbon cross-link and is produced by many streptococci ([5](#)[4](#)–[7](#)).

Installation of the Lys–Trp cross-link in streptide biosynthesis is catalyzed by the radical *S*-adenosylmethionine (SAM) enzyme StrB (5). Homologs SuiB and AgaB carry out similar reactions (8). As in the production of other RiPPs, streptide biosynthesis requires an N-terminal leader sequence preceding the core peptide sequence (5) (Fig. 1A, Top). Although such leader sequences are ubiquitous in RiPP precursors, their roles are still under scrutiny. Mounting evidence suggests that the leader sequence directly interacts with the tailoring enzymes to act as a guide that facilitates the modification (9). Recent crystal structures of the lanthipeptide dehydratase NisB and cyanobactin cyclodehydratase LynD have shed light onto leader-sequence recognition by RiPP-tailoring enzymes (10, 11). In both structures, the leader peptide forms an extended  $\beta$ -sheet with a domain that contains a winged helix-turn-helix (wHTH) topology, which was recently classified by HHPred-based bioinformatics as a RiPP precursor peptide recognition element (RRE) (12). Intriguingly, this study further identified the presence of RRE-like domains in the majority of prokaryotic RiPP classes, including those modified by the so-called SPASM-domain containing radical SAM enzymes. Named after founding members subtilosin A, pyrroloquinoline quinone (PQQ), anaerobic sulfatase, and mycofactocin, this subfamily harbors a Cys-rich C-terminal domain that accommodates binding of one or two additional Fe–S clusters. Notably, StrB, AgaB, and SuiB are also members of this subfamily.

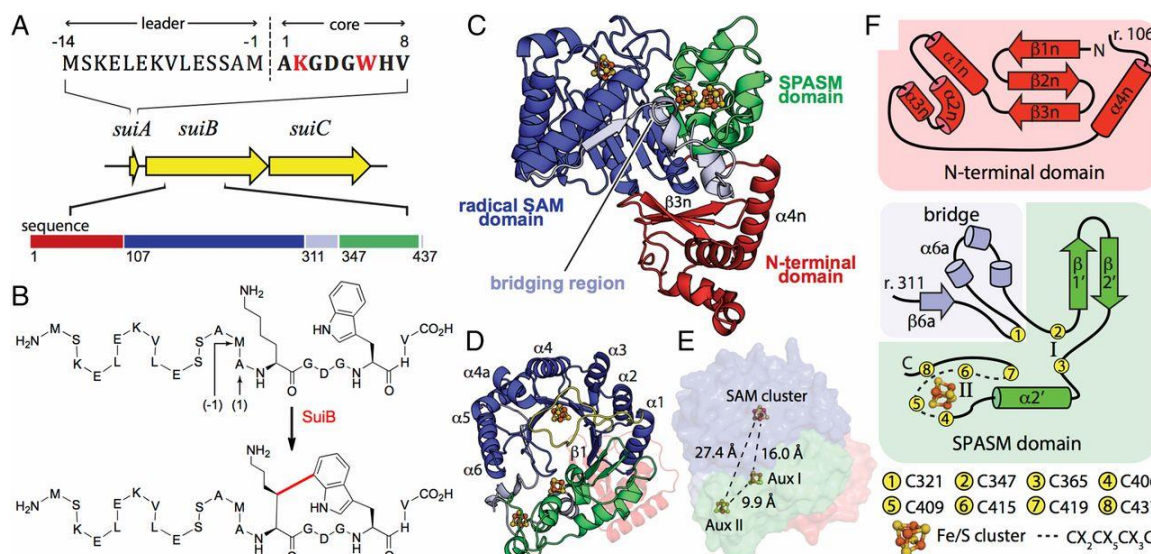


Figure 77-7. Figure 1. The *sui* gene cluster and the reaction catalyzed by SuiB.

(A) The *sui* gene cluster is highly homologous to *str*. It encodes a 22-mer precursor peptide (SuiA), a tailoring radical SAM enzyme (SuiB), and a putative transporter/protease (SuiC). The sequence of SuiA is depicted with the 8-mer sequence of the mature product shown in bold. Cross-linked residues are shown in red. (B) SuiB catalyzes Lys–Trp cross-link formation in SuiA. The new bond installed is shown in red. (C) SuiB contains three [4Fe–4S] clusters and multiple functional domains. (D) The radical SAM domain (blue) forms a partial TIM barrel that is laterally closed by the auxiliary cluster-containing SPASM domain (green). The canonical catalytic [4Fe–4S] cluster-binding motif following  $\beta 1$  is shown in yellow. The bridging region and N-terminal domains are shown in light blue and red, respectively. (E) Placement of the three [4Fe–4S] clusters, shown in ball and stick representation (Fe, orange; S, yellow). Distances are calculated between the nearest atoms. (F) Topologies of the N-terminal domain, SPASM domain, and bridging region.

Radical SAM enzymes reductively cleave SAM bound to a  $[4\text{Fe-4S}]^+$  cluster to generate a 5'-deoxyadenosyl radical (5'-dA $\cdot$ ), which initiates turnover by abstraction of a hydrogen atom from the substrate (13, 14). In the biosynthesis of streptide and its homologs, hydrogen atom abstraction occurs from the lysine  $\beta$ -hydrogen in the precursor peptide (5, 8). The resultant lysyl radical then reacts with the tryptophan side chain to create a cross-link (Fig. 1B). Like other members of the SPASM subfamily, StrB, AgaB, and SuiB, which install Lys–Trp cross-links in their respective peptides, also contain a C-terminal

domain with a characteristic 7-cysteine motif (CX<sub>9-15</sub>GX<sub>4</sub>C-gap-CX<sub>2</sub>CX<sub>5</sub>CX<sub>3</sub>C-gap-C), which allows for binding of additional, so-called "auxiliary" [4Fe-4S] clusters ([15](#)). Although this motif has been shown to be necessary for streptide biosynthesis ([5](#)), the precise role of auxiliary clusters in SPASM enzymes remains an active area of investigation ([16](#)–[18](#)).

Here, we report crystal structures of SuiB and thus a visualization of a RiPP-modifying radical SAM enzyme. Using X-ray diffraction, we determined three structures that illustrate conformational changes associated with binding of SAM and substrate SuiA. These structures depict an N-terminal RRE domain positioned at the entrance to the active site that appears poised to support precursor binding. Surprisingly, however, we detect little interaction between the RRE domain and SuiA and instead observe an  $\alpha$ -helical peptide corresponding to the leader sequence of SuiA bound within the catalytic barrel formed by the radical SAM and SPASM domains. We identify specific hydrogen-bonding interactions made by a region of the barrel that is stabilized by an auxiliary [4Fe-4S] cluster and a highly conserved amino acid motif (LESS) within the SuiA leader sequence. Using computational methods, we further demonstrate that binding of the leader sequence within the catalytic barrel facilitates conformations that position the core sequence within the active site, bringing the cross-linking residues in proximity of the 5'-carbon of 5'-dA. Together, these results provide structural insights into binding of the precursor peptide to SuiB and the conformational changes needed for the unprecedented C-C cyclization reaction.

## 7.4 Results

### 7.4.1 Overall Structure of SuiB.

SuiB is encoded by *Streptococcus suis* and is 96% homologous to the enzyme StrB from *Streptococcus thermophilus*. Both enzymes install Lys–Trp cross-links in their respective substrates, SuiA and StrA ([5](#), [8](#)). Purification and reductive reconstitution of N-terminally His<sub>6</sub>-tagged SuiB leads to an average Fe/S content of  $10.4 \pm 0.1$  Fe and  $9.0 \pm 0.1$  S per protomer. To visualize the overall architecture of SuiB and the conformational changes associated with substrate binding, we determined three crystal structures ([SI Appendix, Tables S1 and S2](#)). A substrate-free crystal structure of SuiB was determined to 2.5-Å resolution. Additionally, crystals were soaked with excess SAM to yield a structure of SAM-bound SuiB to 2.5-Å resolution. The highest resolution structure at 2.1 Å was solved for reconstituted SuiB cocrystallized with excess SAM and precursor peptide SuiA. Crystals did not form in the presence of peptide alone.

The overall structure of SuiB contains three functionally distinct domains ([Fig. 1C](#)), described in detail below: the N-terminal RRE domain (residues 1–106), the radical SAM domain (residues 107–310) followed by a short bridging region (residues 311–346), and the C-terminal SPASM domain (residues 347–437).

### 7.4.2 SuiB Contains a Canonical Radical SAM Domain.

The catalytic core of SuiB (residues 107–439) forms a hollow barrel composed of the radical SAM domain bridged to the C-terminal SPASM domain. Characteristic to many members of the radical SAM superfamily, the SAM domain consists of a partial  $(\beta/\alpha)_6$

triose phosphate isomerase (TIM) barrel ([Fig. 1D](#), blue) that houses the active site [4Fe–4S] cluster in a loop immediately after the  $\beta$ 1 strand ([13](#), [19](#)). Contained within this loop ([Fig. 1D](#), yellow), Cys117, Cys121, and Cys124 form the so-called radical SAM CX<sub>3</sub>CX $\Phi$ C motif (in which  $\Phi$  is an aromatic residue) and ligate three of the four irons in the SAM cluster. As expected, our substrate-free structure has an open coordination site at the remaining "unique Fe," whereas in our SAM-soaked structure, we observe intact SAM forming a chelate at this position ([SI Appendix, Fig. S1](#)). Additional structural motifs critical for SAM binding are also conserved in SuiB ([SI Appendix](#)), such as hydrogen bonding between the main-chain carbonyl oxygen of the hydrophobic residue  $\Phi$  (Phe123) and the N6-amino group of adenine ([19](#)) ([SI Appendix, Fig. S1](#)). Although a canonical TIM barrel is composed of eight strands and eight helices, the entire fold has rarely been observed among radical SAM enzymes ([13](#), [19](#)). Instead, the C-terminal SPASM domain provides a lateral closure in SuiB ([16](#), [17](#), [20](#)) ([Fig. 1D](#), green).

#### 7.4.3 The Bridging Region Provides a Critical Residue for Auxiliary Cluster Ligation.

Previous biochemical analyses with StrB and SuiB have not unambiguously determined the number of auxiliary clusters. Our structures clearly reveal two intact auxiliary [4Fe–4S] clusters ([Fig. 1E](#)), similar to anSMEcpe, the anaerobic sulfatase maturing enzyme from *Clostridium perfringens* and the only other structurally characterized member of the SPASM subfamily ([17](#)). Full ligation of two auxiliary clusters in SuiB is enabled by an eighth anterior cysteine (Cys321) within the bridging region. This cysteine is located much farther upstream compared with previously characterized SPASM enzymes, at a position 26 residues before the 7-cysteine motif ([Fig. 1F](#), light blue). Following the C

terminus of the radical SAM domain, the  $\beta$ 6a strand of the bridging region dips into the barrel to provide this initial coordinating residue for the first auxiliary cluster, Aux I, before exiting the barrel as a fragmented helix ([Fig. 1 C and F](#)). Although the presence of a cysteine within the bridging region appears to be common to many SPASM radical SAM enzymes ([15](#)–[17](#), [20](#), [21](#)), the large gap in sequence appears to be unique to characterized members within this subfamily and a key feature of Lys–Trp cross-linking enzymes ([5](#)).

#### 7.4.4 The SPASM Domain Binds Two Fe–S Clusters.

The SPASM domain incorporates the auxiliary clusters around a  $\beta$ -hairpin and  $\alpha$ -helix. It is initiated by two cysteines, Cys347 and Cys365, which, in addition to Cys-321, ligate Aux I and flank the  $\beta$ -hairpin. The first strand of the SPASM motif ( $\beta$ 1') interacts with  $\beta$ 1 of the radical SAM domain to extend the  $\beta$ -sheet within the barrel ([Fig. 1 D and F](#)). Following residue Cys365, the 7-cysteine SPASM motif is punctuated by the  $\alpha$ 2' helix and resumes with the remaining CX<sub>2</sub>CX<sub>5</sub>CX<sub>3</sub>C sequence. The first three cysteines in this sequence—Cys406, Cys409, and Cys415—encircle and ligate the second auxiliary cluster, Aux II, before the chain hooks back to provide the final coordinating residue for Aux I, Cys419. The SuiB sequence then doubles back again to fully coordinate Aux II with the final cysteine of the SPASM motif, Cys437, after which it terminates at Leu439, in contrast to the extended C-terminal helix ( $\alpha$ 6') observed in anSMEcpe ([17](#)) ([Fig. 1F](#) and [SI Appendix, Fig. S2](#)). Aux I and II are located 16.0 Å and 27.4 Å from the radical SAM cluster, respectively, measured between the closest atoms ([Fig. 1E](#)). As in anSMEs,

full cysteine ligation precludes a substrate-binding role for the auxiliary clusters, a conclusion that is further supported by the SuiA-bound structure described below.

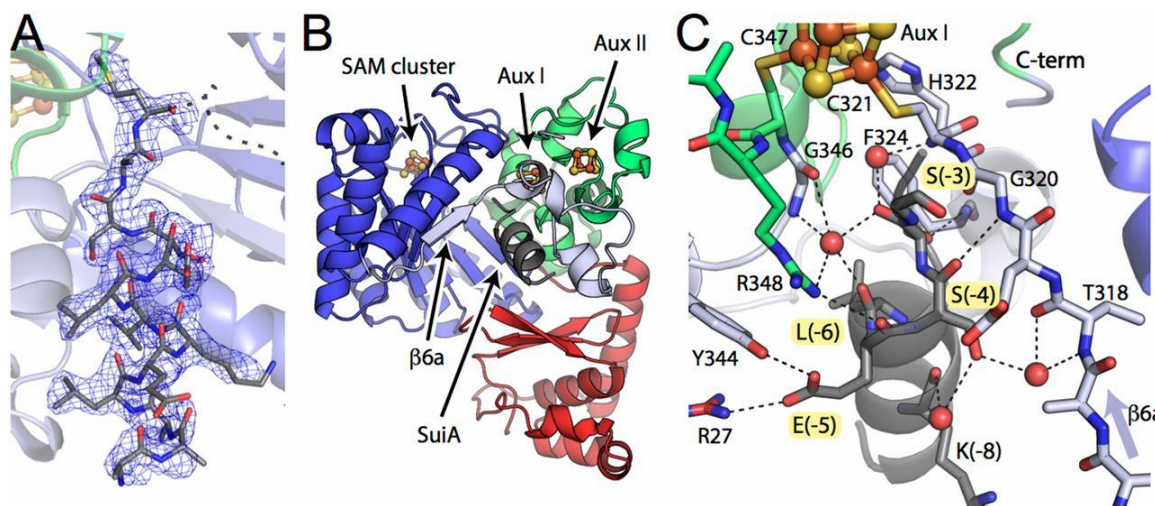
#### 7.4.5 The N-Terminal Domain Adopts an RRE Fold.

As predicted by previous bioinformatics studies ([12](#)), the N-terminal domain of SuiB shows structural homology with the archetypal RRE domain, PqqD ([22](#), [23](#)). Although PqqD is believed to act as a peptide chaperone for PqqE, the radical SAM enzyme involved in PQQ biosynthesis, these two proteins have never been visualized in complex ([24](#)–[26](#)). Our structures thus provide insight into the arrangement of an RRE-like domain associated with a radical SAM enzyme. The RRE domain in SuiB is initiated by a three-stranded antiparallel  $\beta$ -sheet that adjoins a trihelical bundle ([Fig. 1F](#), red), forming a wHTH-like motif that protrudes from the catalytic core ([27](#)) ([Fig. 1C](#), red). This domain is then anchored in a cleft formed between the SPASM domain ([Fig. 1C](#), green) and adjacent bridging region ([Fig. 1C](#), light blue) via an additional helix  $\alpha 4_n$  ([Fig. 1C](#) and [SI Appendix, Fig. S3](#)), placing the  $\beta$ -sheet of the wHTH motif above the TIM barrel entrance. In the recent structures of NisB and LynD, the leader peptides are observed in between  $\alpha 3_n$  and  $\beta 3_n$ , forming an extended antiparallel  $\beta$ -sheet with the wHTH motif ([10](#), [11](#)) ([SI Appendix, Fig. S3](#)). Differences in sequence between SuiB and its close homologs are also concentrated in this groove ([SI Appendix, Fig. S4](#)).

#### 7.4.6 Recognition of the Leader Sequence by the Catalytic Core.

The enzyme crystallized in the presence of SAM and SuiA yields clear density for the leader portion of the substrate peptide (residues  $-13$  to  $-1$ ) ([Fig. 2A](#)), while density for

the core sequence (residues 1–8) is disjointed and difficult to assign. In contrast to the structures of NisB and LynD, where the RRE domains make many direct interactions with the respective substrates ([10](#), [11](#)) ([SI Appendix, Fig. S3](#)), we observe the leader sequence of SuiA bound within the SuiB barrel, adjacent to both the bridging region, which provides the first cysteine ligand for Aux I, and the SPASM domain ([Fig. 2B](#)). Furthermore, the  $\alpha$ -helical nature of the SuiA leader, predicted from sequence analysis, is maintained within the catalytic barrel before it transitions into a loop, whose contiguous density terminates immediately before the core sequence, adjacent to Aux I ([Fig. 2C](#)). While many RiPP leader peptides have been shown to adopt  $\alpha$ -helical conformations in trifluoroethanol, the persistence of this secondary structure upon binding to the tailoring enzyme has only been observed in our structures and the recent structure of MdnC, which binds the leader peptide as a single-turn  $\alpha$ -helix but lacks sequence homology to a typical RRE ([3](#), [9](#), [11](#), [28](#)).



**Figure 78-7.** Figure 2. SuiA recognition in the active site is dominated by interactions of the leader sequence with the bridging region.

(A) Observed electron density for the helical leader sequence displayed as a 2F<sub>o</sub>-F<sub>c</sub> composite omit map contoured at 1.0  $\sigma$ . The disordered core sequence is represented by a dashed line. (B) SuiA (dark gray) binds in the catalytic barrel rather than to the putative recognition element, the N-terminal domain (red). (C) Hydrogen-bonding network of SuiA (dark gray) bound in the active site adjacent to the bridging region (light blue) and SPASM domain (green). Ordered water molecules are shown as red spheres. The LESS motif is highlighted. See SI Appendix, Fig. S5A for a stereoview.

Recognition of the SuiA leader sequence is primarily achieved through interactions with the bridging region ([Fig. 2C](#)). These interactions orient the substrate helix and thereby facilitate proper arrangement of the core sequence in the active site ([Fig. 2C](#) and [SI Appendix, Fig. S5](#)). Perhaps explaining its high conservation in streptide precursor peptides ([5](#)), the LESS motif of the leader sequence (residues -6 to -3) plays a particularly important role in orienting SuiA by providing the only hydrogen-bonding partners with SuiB. This hydrogen-bonding network is initiated by water-mediated interactions between SuiA-Leu(-6) and Gly346 in the bridging region and by Arg348 in the SPASM domain ([Fig. 2C](#) and [SI Appendix, Fig. S5](#)). The SPASM domain further hydrogen bonds with the backbone carbonyl oxygens of both SuiA-Glu(-5) and, indirectly, SuiA-Ser(-3) through Arg348. The only observed interaction with the RRE motif is made between SuiA-Glu(-5) and Arg27 in the form of a salt bridge. The remaining interactions occur with the bridging region and include Tyr344/SuiA-Glu(-5), Gly320/SuiA-Ser(-4), Phe324/SuiA-Ser(-3), and indirect water-mediated interactions with Thr318, His322, and Gly346 ([Fig. 2C](#) and [SI Appendix, Fig. S5](#)). The buried surface area along the peptide-protein interface spanning from the barrel opening to the active site is 730 Å<sup>2</sup>, almost 51% of the modeled peptide's total surface area.

#### 7.4.7 Substrate Binding Is Coupled to Loop Movements.

The three snapshots obtained in this study illustrate the conformational changes associated with substrate binding by SuiB. Comparison of the substrate-free and SAM-soaked structures shows minimal conformational changes associated with binding of SAM alone (average C $\alpha$  rmsd of  $0.285 \pm 0.037$  Å) ([Fig. 3A](#), gray curve). In the absence of peptide, we observe density for intact SAM bound in the active site, suggestive of a preturnover state ([SI Appendix, Fig. S1](#)). Crystallization with SAM and SuiA results in large-scale rearrangements ([Fig. 3A](#), red/blue curves). At the bottom of the barrel, as oriented in [Fig. 1C](#), the largest changes are seen in the RRE-like domain and the  $\alpha 6$  helix, which are angled farther away from the barrel opening upon binding of SuiA ([Fig. 3A](#) and [SI Appendix, Fig. S6 A and B](#)). These regions make a number of crystal lattice contacts, making further interpretation difficult; however, it is evident that the enzyme can accommodate significant motions, particularly in the RRE domain. In contrast, motion at the top of the barrel is unencumbered by crystal contacts and is dominated by two loops, L1 and L2, linked by hydrogen-bonding interactions between the backbone carbonyl of Gly122 and backbone amides of Lys286 and Ile287 ([Fig. 3B](#) and [SI Appendix, Fig. S6C](#)). In particular, residues 125–134 of L1 and 279–285 of L2 adopt a new conformation in the SuiA-bound structure, which occludes the active site. Interestingly, L1 directly follows the SAM-cluster binding motif, and its displacement is likely a result of SAM cleavage in the active site.

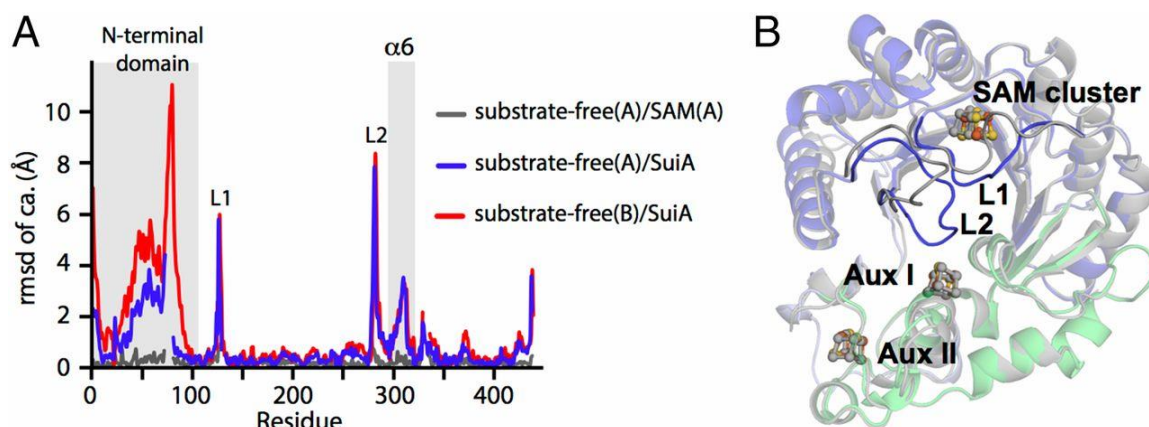


Figure 79-7. Figure 3. Substrate binding leads to coordinated loop movements.

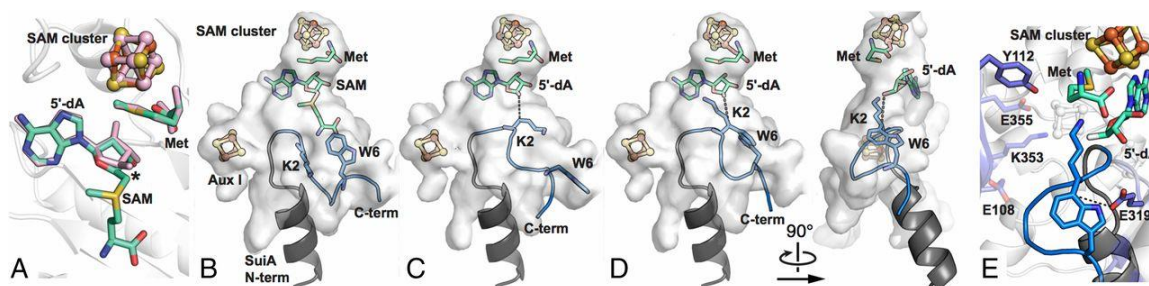
(A) Rmsd of the C $\alpha$  atoms versus residue number upon SAM and SAM + SuiA binding, respectively. The two chains (A/B) in the asymmetric unit are denoted parenthetically. Binding only SAM leads to minimal changes (gray curve), yielding an average C $\alpha$  rmsd of  $0.285 \pm 0.037$  Å, whereas additional binding of SuiA leads to greater changes (blue/red), particularly in the loops L1 and L2. See SI Appendix, Fig. S6 for additional views. (B) Visualization of loop motions upon SuiA binding. The substrate-free enzyme is shown in gray. The RRE is omitted for clarity.

#### 7.4.8 A Postturnover Conformation in SuiA-Bound Structure.

In our SuiA-bound structure, we observe two disconnected regions of density in close proximity to the SAM cluster ([SI Appendix, Fig. S1](#)). The first supports Met bound as a tridentate chelate to the unique Fe through its  $\alpha$ -amine, carboxylate, and side-chain sulfur atom, an arrangement that mirrors previously observed post-SAM cleavage structures ([29–31](#)). This is consistent with prereduction of SuiB with sodium dithionite before crystallization ([8](#)). Surprisingly, however, 5'-dA cannot completely account for the second region of density. Instead, we observe intact SAM at this position ([SI Appendix, Fig. S1](#)). In the absence of a suitable reductant, this unusual feature possibly occurred following a single turnover, in which dynamic motion of the enzyme trapped excess

SAM in the active site, perhaps poised to replace the Met bound on the SAM cluster in preparation for another catalytic cycle.

The active-site hydrogen-bonding arrangement persists in the SuiA-bound and SAM-bound structures, with additional hydrogen bonds observed from Glu319 and Arg272 to the methionine moiety of the trapped SAM (13, 19) (*SI Appendix, Fig. S1*). Further analysis of the 5'-dA portion of SAM shows that the 5'-C of the ribose group tilts down toward the peptide substrate, priming the enzyme for H-atom abstraction. This arrangement of the 5'-dA moiety of SAM mimics previously observed SAM-cleavage products in other radical SAM enzymes (29–31) (*Fig. 4A*). Together, these observations suggest that SuiB in complex with SuiA is trapped in a post-SAM cleavage conformation.



**Figure 80-7. Figure 4.** Binding of the SuiA leader sequence supports positioning of the core sequence in the active site of SuiB.

(A) The arrangement of Met and SAM (green) in our SuiA-bound structure mimics previously observed SAM-cleavage products from RlmN (pink) (29–31). The 5'-C of 5'-dA is marked with an asterisk. (B–D) The active-site cavity of SuiB is shown as a white surface. The crystallographic model of the leader peptide is shown in gray. Rosetta-based simulations yield low-energy conformations of the core sequence (blue) within our SuiA-bound crystal structure both with (B) intact SAM and (C) the methionine moiety of SAM removed. (D) Simulations also yield low-energy conformations of the cyclized core peptide (blue) within the active-site cavity. (E) Of the five titratable residues (shown as sticks) near the active site, simulations favor E319 (~4.9 Å from the SuiA-W6 C7 position) as the catalytic base.

Consistent with this post-SAM cleavage model are the shifts observed in L1 upon SuiA binding. The adenine and ribose moieties of SAM make all of the previously observed contacts observed for the 5'-dA moiety in the SAM-only structure. As in the SPASM-containing enzymes anSMEcpe ([17](#)) and sporulation-killing factor maturase SkfB ([32](#)), as well as the partial-SPASM enzyme 2-deoxy-scylo-inosamine dehydrogenase BtrN ([16](#)), an additional aromatic residue directly follows the final cysteine of the radical SAM motif. In SuiB, this residue is Phe125 ([SI Appendix, Fig. S2](#)). While the position of the adenine moiety is virtually unchanged between structures, L1 shifts to create an additional hydrogen bond between the backbone of Phe125 and the N7 position of adenine ([SI Appendix, Fig. S1](#)). Concurrently, the Phe125 side chain flips to stack perpendicularly with the nucleobase, creating a hydrophobic pocket that likely would facilitate stabilization of the cleaved products ([SI Appendix, Fig. S1F](#)).

Coupled together through hydrogen-bonding interactions, changes in L1 result in coordinated rearrangement of L2. Furthermore, movement of Phe125 generates an additional interaction between its backbone carbonyl group and the side chain of Thr282 (L2) ([SI Appendix, Fig. S6C](#)). As a result, the residues of L2 move together to cap the barrel ([Fig. 3B](#)), perhaps also inducing significant motion in the downstream  $\alpha 6$  helix. A comparison of surface renderings, with and without peptide bound, visually confirms constriction of the channel supplying bulk-solvent access to the active site from the top of the barrel ([SI Appendix, Fig. S6B](#)). It is possible that, by limiting solvent access, L2 facilitates a dielectric change in the active site, which has previously been proposed to lower the free-energy barrier for SAM cleavage ([33](#)). Intriguingly, an associated channel

also connects Aux I to SuiA, suggesting that loop motions may be important for redox reactions involving Aux I.

#### 7.4.9 The Leader Sequence Helps Position the Core Sequence.

The final eight residues of SuiA are disordered in our crystal structure. To investigate the possible conformations adopted by these eight residues within the SuiA-bound crystal structure, two Rosetta-based simulations were performed: one with the Lys–Trp-cyclized SuiA and a second with the linear SuiA substrate. First, the cyclized SuiA core peptide was modeled using NMR-derived constraints for streptide (5) and placed in the active site of the SuiA-bound structure using geometric constraints imposed by a peptide bond linkage to the crystallographic model of the leader peptide. This simulation led to SuiA placements showing considerable steric clashes with the methionine moiety of SAM, which occupies the 5'-dA site. We next modeled uncyclized SuiA in the active site of our SuiA-bound structure. This simulation yielded multiple low-energy conformations of SuiA compatible with the crystallographic model of SuiB (Fig. 4B, modeled region of SuiA shown in blue) displaying conformational heterogeneity (C $\alpha$  rmsds of up to 3.4 Å) that may explain the observed lack of resolvable electron density in our crystal structure.

As the adenine and ribose moieties of SAM in our SuiA-bound structure mimic the postcleavage conformation of 5'-dA (29↓–31) (Fig. 4A), two additional simulations were performed to model the final eight SuiA residues within our SuiA-bound structure with the methionine moiety of SAM removed. Eleven residues lining the barrel, including those that form hydrogen bonds with the methionine moiety of SAM, were allowed to sample other side-chain conformations (SI Appendix). C $\beta$  of SuiA-K2 was additionally

constrained to within reasonable hydrogen abstraction distances (2.8–4.3 Å) from the adenosine 5'-C ([Fig. 4 C and D](#), dotted line). Modeling uncyclized SuiA led to multiple low-energy conformations ([Fig. 4C](#), modeled region of SuiA shown in blue), suggesting that the leader sequence can facilitate correct positioning of the substrate. In the final simulation, cyclized SuiA was modeled into the active site ([Fig. 4D](#), modeled region of SuiA shown in blue). Four clusters of low-energy conformations were observed, displaying relatively high backbone similarity ([SI Appendix, Figs. S7 and S8](#)). Of the 11 SuiB residues allowed to sample other side-chain conformations, Gln26, Glu108, and Asn315 adopt a conformation not seen in the crystal structure, while Glu319 and Arg272 adopt multiple conformations, including those seen in the structure. Notably, Glu319 and Arg272, which interact with the methionine moiety of SAM in the SuiA-bound structure ([SI Appendix, Fig. S1D](#)), form new interactions with the substrate peptide in these simulated models. In all of these conformations, Glu319 is the closest residue to C7 of SuiA-W6 (within ~5 Å) ([Fig. 4E](#)). Overall, these simulations are consistent with a scenario in which the SuiA leader sequence positions the core sequence into the active site of SuiB for posttranslational modification.

## 7.5 Discussion

In addition to exploring the conformational changes that facilitate Lys–Trp cross-link formation, the crystal structures presented here provide insights into the functions of ancillary domains that are prevalent in the radical SAM enzyme superfamily. As expected from bioinformatic analyses, the N-terminal domain in SuiB adopts a RRE fold that is docked at the opening of the radical SAM catalytic core, poised to mediate peptide

delivery. However, rather than binding to the RRE domain, we observe the SuiA leader peptide primarily interacting with the catalytic barrel. Unique among published RRE-containing structures, this discovery not only elucidates leader peptide function but also provides insights into the role of the RRE domain during catalysis. Simulations of the core sequence further support the role of the leader peptide in guiding posttranslational modifications, while the observed interactions between SuiA and SuiB highlight the importance of a bridging loop linking the radical SAM and SPASM domains. As only the second crystallographically characterized radical SAM enzyme to contain three [4Fe–4S] clusters, SuiB provides additional insights into the SPASM domain and the RRE domain and highlights the unsuspected importance of the bridging domain during catalysis ([34](#)).

While more than a third of SPASM-containing enzymes include a cysteine in the bridging region ([35](#)), the prevalence of fully ligated auxiliary clusters is unknown. Coordination of the upstream cysteine in SuiB precludes direct substrate binding and establishes this feature, first observed in anSMEcpe, as a significant auxiliary cluster-binding motif. There are, however, critical differences between SuiB and anSMEcpe, including the remote position of the upstream cysteine and fragmentation of the  $\alpha_6$  helix ([SI Appendix, Fig. S9](#)). In addition, the arrangement of the bridging region is indicative of a structural role for Aux I that is further supported by mutagenesis studies, in which C347A/C365A mutants are recalcitrant to purification ([8](#)). This region further provides critical contacts for both SuiA and the peptidyl substrate surrogates of anSMEcpe and likely serves a similar binding role in other SPASM-containing enzymes. Thus, variability in the bridging region from one enzyme to another may indicate adaption to the cognate peptide substrate.

Although structural insights remain scarce ([10](#), [11](#), [23](#), [36](#), [37](#)), recent bioinformatic identification of an RRE motif across all RiPP classes has provided invaluable clues to understanding interactions between RiPP precursor peptides and tailoring enzymes. Supported by structural and biochemical analyses, this domain has been implicated in peptide recognition and recruitment. Intriguingly, the structures of SuiB provide an example of a precursor binding location distinct from the RRE domain. This unique SuiA-binding mode suggests that the RRE-like domain in SuiB is either vestigial or involved in an undetected interaction. Observed motion of the N-terminal domain appears to support the latter, and one can envision a simple scenario in which the RRE both recognizes the peptide and delivers it to the active site but at a certain stage in the catalytic cycle releases the precursor peptide. Recent biochemical analysis of the PqqD/PqqE system not only detected peptide binding to the canonical RRE but also confirmed interaction between the peptide chaperone and radical SAM enzyme. Perhaps the RRE domain in SuiB serves as a similar intermediate binding site ([12](#)).

The observed location of SuiA within the barrel may provide insights into mechanism of SuiB. In the catalytic scheme proposed, an active site base facilitates rearomatization by deprotonating the putative tryptophanyl radical intermediate followed by electron transfer to an auxiliary cluster ([5](#)). Inspection of the barrel interior yields five possible titratable side chains within  $\sim 10$  Å of the adenosine 5'-carbon: Glu319, Glu108, Glu355, Tyr112, and Lys353 ([Fig. 4E](#)). The positions of simulated Lys–Trp cross-link conformations, supported by alignments with substrate-bound structures of anSMEcpe and RlmN ([SI Appendix, Fig. S10](#)), favor Glu319 for direct active-site deprotonation ([Fig. 4E](#)). In this step of the catalytic cycle, Aux II is clearly an unsuitable direct electron acceptor, as it is

too far removed from the active site. Electron transfer likely proceeds to Aux I first. Decreased access to bulk solvent as a function of loop movements could then justify electron transfer from Aux I to Aux II and then to a protein redox partner (38).

In conclusion, we present a sequence of structures that not only helps to elucidate the formation of a streptide C–C cross-link but also provides insights into the interplay between RiPP precursor peptides and tailoring enzymes more generally. The structures presented here further demonstrate that the mode of substrate binding greatly contributes to structural diversity within ancillary domains of the radical SAM superfamily. In particular, we gain a newfound appreciation for the bridging region between the SAM and SPASM domains. It will be fascinating to see if future investigations into peptide recognition and recruitment by RiPP-modifying enzymes uncover similar interactions between distinct RRE and catalytic domains, especially those involving radical SAM enzymes like the PqqD/PqqE system.

## 7.6 References

1. Ortega MA, van der Donk WA: **New insights into the biosynthetic logic of ribosomally synthesized and post-translationally modified peptide natural products.** *Cell Chem Biol* 2016, **23**:31–44.
2. Felnagle EA, et al.: **Nonribosomal peptide synthetases involved in the production of medically relevant natural products.** *Mol Pharm* 2008, **5**:191–211.

3. Arnison PG, et al: **Ribosomally synthesized and post-translationally modified peptide natural products: Overview and recommendations for a universal nomenclature.** *Nat Prod Rep* 2013, **30**:108–160.
4. Finking R, Marahiel MA: **Biosynthesis of nonribosomal peptides1.** *Annu Rev Microbiol* 2004, **58**:453–488.
5. Schramma KR, Bushin LB, Seyedsayamdost MR: **Structure and biosynthesis of a macrocyclic peptide containing an unprecedented lysine-to-tryptophan crosslink.** *Nat Chem* 2015, **7**:431–437.
6. Fleuchot B, et al.: **Rgg proteins associated with internalized small hydrophobic peptides: A new quorum-sensing mechanism in streptococci.** *Mol Microbiol* 2011, **80**:1102–1119.
7. Ibrahim M, et al.: **Control of the transcription of a short gene encoding a cyclic peptide in *Streptococcus thermophilus*: A new quorum-sensing system?** *J Bacteriol* 2007, **189**:8844–8854.
8. Schramma KR, Seyedsayamdost MR: **Lysine-tryptophan-crosslinked peptides produced by radical SAM enzymes in pathogenic streptococci.** *ACS Chem Biol* 2017, **12**:922–927.
9. Oman TJ, van der Donk WA: **Follow the leader: The use of leader peptides to guide natural product biosynthesis.** *Nat Chem Biol* 2010, **6**:9–18.

10. Koehnke J, et al.: **Structural analysis of leader peptide binding enables leader-free cyanobactin processing.** *Nat Chem Biol* 2015, **11**:558–563.
11. Ortega MA, et al.: **Structure and mechanism of the tRNA-dependent lantibiotic dehydratase NisB.** *Nature* 2015, **517**:509–512.
12. Burkhart BJ, Hudson GA, Dunbar KL, Mitchell DA: **A prevalent peptide-binding domain guides ribosomal natural product biosynthesis.** *Nat Chem Biol* 2015, **11**:564–570.
13. Broderick JB, Duffus BR, Duschene KS, Shepard EM: **Radical S-adenosylmethionine enzymes.** *Chem Rev* 2014, **114**:4229–4317.
14. Frey PA, Booker SJ: **Radical mechanisms of S-adenosylmethionine-dependent enzymes.** *Adv Protein Chem* 2001:1–45.
15. Haft DH, Basu MK: **Biological systems discovery in silico: Radical S-adenosylmethionine protein families and their target peptides for posttranslational modification.** *J Bacteriol* 2011, **193**:2745–2755.
16. Goldman PJ, Grove TL, Booker SJ, Drennan CL: **X-ray analysis of butirosin biosynthetic enzyme BtrN redefines structural motifs for AdoMet radical chemistry.** *Proc Natl Acad Sci USA* 2013, **110**:15949–15954.
17. Goldman PJ, et al.: **X-ray structure of an AdoMet radical activase reveals an anaerobic solution for formylglycine posttranslational modification.** *Proc Natl Acad Sci USA* 2013, **110**:8519–8524.

18. Grove TL, Lee K-H, St Clair J, Krebs C, Booker SJ: **In vitro characterization of AtsB, a radical SAM formylglycine-generating enzyme that contains three [4Fe-4S] clusters.** *Biochemistry* 2008, **47**:7523–7538.
19. Dowling DP, Vey JL, Croft AK, Drennan CL: **Structural diversity in the AdoMet radical enzyme superfamily.** *Biochim Biophys Acta* 2012, **1824**:1178–1195.
20. Hänzelmann P, Schindelin H: **Crystal structure of the S-adenosylmethionine-dependent enzyme MoaA and its implications for molybdenum cofactor deficiency in humans.** *Proc Natl Acad Sci USA* 2004, **101**:12870–12875.
21. Haft DH: **Bioinformatic evidence for a widely distributed, ribosomally produced electron carrier precursor, its maturation proteins, and its nicotinoprotein redox partners.** *BMC Genomics* 2011, **12**:21.
22. Söding J: **Protein homology detection by HMM-HMM comparison.** *Bioinformatics* 2005, **21**:951–960.
23. Tsai T-Y, Yang C-Y, Shih H-L, Wang AHJ, Chou S-H: ***Xanthomonas campestris* PqqD in the pyrroloquinoline quinone biosynthesis operon adopts a novel saddle-like fold that possibly serves as a PQQ carrier.** *Proteins Struct Funct Bioinf* 2009, **76**:1042–1048.
24. Weeksler SR, et al.: **Interaction of PqqE and PqqD in the pyrroloquinoline quinone (PQQ) biosynthetic pathway links PqqD to the radical SAM superfamily.** *Chem Commun (Camb)* 2010, **46**:7031–7033.

25. Latham JA, Iavarone AT, Barr I, Juthani PV, Klinman JP: **PqqD is a novel peptide chaperone that forms a ternary complex with the radical S-adenosylmethionine protein PqqE in the pyrroloquinoline quinone biosynthetic pathway.** *J Biol Chem* 2015, **290**:12908–12918.
26. Barr I, et al.: **The pyrroloquinoline quinone (PQQ) biosynthetic pathway: Demonstration of de novo carbon-carbon cross-linking within the peptide substrate (PqqA) in the presence of the radical SAM enzyme (PqqE) and its peptide chaperone (PqqD).** *J Biol Chem* 2016, **291**:8877–8884.
27. Brennan RG: **The winged-helix DNA-binding motif: Another helix-turn-helix takeoff.** *Cell* 1993, **74**:773–776.
28. Li K, Condurso HL, Li G, Ding Y, Bruner SD: **Structural basis for precursor protein-directed ribosomal peptide macrocyclization.** *Nat Chem Biol* 2016, 12:973–979.
29. Hänzelmann P, Schindelin H: **Binding of 5'-GTP to the C-terminal FeS cluster of the radical S-adenosylmethionine enzyme MoaA provides insights into its mechanism.** *Proc Natl Acad Sci USA* 2006, **103**:6829–6834.
30. Schwalm EL, Grove TL, Booker SJ, Boal AK: **Crystallographic capture of a radical S-adenosylmethionine enzyme in the act of modifying tRNA.** *Science* 2016, **352**:309–312.

31. Rohac R, et al.: **Carbon-sulfur bond-forming reaction catalysed by the radical SAM enzyme HydE.** *Nat Chem* 2016, **8**:491–500.
32. Bruender NA, Bandarian V: **SkfB abstracts a hydrogen atom from C $\alpha$  on SkfA to initiate thioether cross-link formation.** *Biochemistry* 2016, **55**:4131–4134.
33. Shisler KA, Broderick JB: **Emerging themes in radical SAM chemistry.** *Curr Opin Struct Biol* 2012, **22**:701–710.
34. Grell TAJ, Goldman PJ, Drennan C: **SPASM and twitch domains in S-adenosylmethionine (SAM) radical enzymes.** *J Biol Chem* 2015, **290**:3964–3971.
35. Akiva E, et al: **The structure–function linkage database.** *Nucleic Acids Res* 2014, **42**:D521–D530.
36. Ortega MA, et al.: **Structure and tRNA specificity of MibB, a lantibiotic dehydratase from actinobacteria involved in NAI-107 biosynthesis.** *Cell Chem Biol* 2016, **23**:370–380.
37. Regni CA, et al.: **How the MccB bacterial ancestor of ubiquitin E1 initiates biosynthesis of the microcin C7 antibiotic.** *EMBO J* 2009, **28**:1953–1964.
38. Moser CC, Anderson JLR, Dutton PL: **Guidelines for tunneling in enzymes.** *Biochim Biophys Acta* 2010, **1797**:1573–1586.

## 7.7 SI Results and Discussion (cont.)

### 7.7.1 SAM binding and the Active Site

Hydrogen bond stabilization of the various SAM moieties is critical for positioning the cosubstrate about the cluster (Fig. S1). Related hydrogen-bonding motifs include the "GGE" motif, important for methionine orientation; the ribose motif; as well as the "GXIXGXXE" motif and  $\beta 6$  or "adenine binding" motif, both involved in stabilizing the adenine moiety (1-3). Interactions with the adenine component include hydrogen bonds to the Ser279 backbone of the  $\beta 6$  strand, the Phe123 carbonyl group, as well as hydrophobic interactions with Val249. Likewise, side chains of Ser210 and Asn247, residues in the ribose and GXIXGXXE motifs respectively, help to position the ribose moiety by providing hydrogen-bonding partners for the 3'-hydroxyl group. It is clear from the higher resolution post-cleavage structure that Gln212 also indirectly contributes to ribose stabilization through a water-mediated hydrogen bond. In contrast, proper arrangement of methionine is primarily provided by hydrogen bonds with the GGE motif, following the  $\beta 2$  strand. To facilitate bonding, the carbonyl group of the second glycine points toward methionine, making it a rare cis-isomer. While common to many radical SAMs, this Gly160-Met interaction is supplemented by additional hydrogen bonding with the Glu161 side chain.

### 7.7.2 Evaluation of large-scale movements upon substrate binding

An examination of the post SAM-cleavage structure reveals significant conformational changes in both the radical SAM and RRE domains (Fig. S6A). Although minimal interactions are observed between SuiA and the RRE in SuiB, calculations comparing the

N-terminal domains yield a C $\alpha$  RMSD of 1.85 Å or 3.97 Å for chains A and B respectively, in contrast to an average of only  $0.294 \pm 0.011$  Å upon SAM binding (Fig. 4A). The wHTH domain, in particular, appears to angle outward away from the mouth of the barrel. The intensity of this effect is chain dependent, as is the involvement of  $\alpha 4n$  (Fig. 4A and S6A).

### 7.7.3 Additional helix links wHTH domain to the catalytic barrel

An analogous helix to  $\alpha 4n$  has been observed in the microcin C biosynthetic enzyme MccB for which the RRE serves as a peptide clamp (4). Structurally, this helix follows consecutively from the wHTH domain in SuiB, while the MccB homodimer utilizes a domain-swapping mechanism to generate a similar motif. In both enzymes, the ancillary helix (Fig. S3, grey) forms significant van der Waals and hydrophobic interactions with the primary helix,  $\alpha 1$ , of the adjacent bundle at an approximate crossing angle of 50°. Although direct comparisons are difficult given the precursor peptide of MccB lacks a leader sequence, this additional helix may play a structural role in orienting the RRE with respect to the catalytic core. Unfortunately, without peptide bound to the RRE, it is unclear whether it is required for peptide binding, or simply serves as a stabilizing link to the catalytic domain.

## 7.8 SI Materials and Methods

### 7.8.1 Materials and Strains

The genomic DNA of *Streptococcus suis* 92-4172 was kindly provided by Prof. Marcelo Gottschalk at the University of Montreal, Canada. SuiB was cloned as a hexa-His-tagged

construct, purified, reconstituted, and pre-reduced with sodium dithionite as recently reported (5). Reconstituted SuiB contained  $10.4 \pm 0.1$  Fe and  $9.0 \pm 0.1$  S per protomer. SuiA was synthesized and purified as described (5). Its identity was verified by high-resolution (HR) HPLC-MS ( $[M+2H]^{2+}$  calc 1216.6016  $[M+2H]^{2+}$  obs 1216.60301,  $\Delta p p m \sim 1.2$ ). Wt, reconstituted SuiB turned over substrate SuiA with a  $V_{max}/[E]_T$  of 0.18 min<sup>-1</sup>.

### 7.8.2 Crystallization

Crystals of N-terminally His6-tagged SuiB were grown anaerobically in a glove box (Coy Laboratory Products) under a 97% N<sub>2</sub>, 3% H<sub>2</sub> atmosphere using the sitting well vapor diffusion method. Crystallization trays were chilled on a cold block ( $\sim 4^\circ\text{C}$ ) during preparation, and all solutions were incubated at  $12^\circ\text{C}$  (Torrey Pines Scientific Incubator) prior to mixing to minimize nucleation events. All trays were incubated and maintained at  $12^\circ\text{C}$  during growth and storage. To obtain the SuiA-bound structure, a solution containing 23 mg/mL of His6-tagged SuiB in storage buffer [100 mM HEPES, pH 7.5, 300 mM KCl, 5 mM DTT, 10% (v/v) glycerol] was mixed with a stock solution of SuiA in storage buffer lacking DTT, yielding a final SuiA concentration of 1.9 mM. The resulting solution was incubated at  $12^\circ\text{C}$  for 10 min after which it was combined 1:1 with precipitant solution to form a 4  $\mu\text{L}$  drop. The precipitant solution was generated by combining 100 mM Bis-Tris, pH 6.0, 200 mM Li<sub>2</sub>SO<sub>4</sub>, 27% (w/v) PEG 3350 with 210 mM SAM in water to yield a final SAM concentration of 10.5 mM. Crystals appeared within 2 days and were fully formed ( $\sim 75 \times 75 \mu\text{m}^2$ ) within a week. Sheet-like crystals were gently separated, looped and transferred briefly into cryoprotectant [200 mM

Li<sub>2</sub>SO<sub>4</sub>, 53.8 mM BISTRIS, 27% (w/v) PEG 3350, 26% (v/v) PEG 400, 6 mM SAM] before cryocooling in liquid nitrogen. To obtain the apo and SAM-bound structures, a solution containing 18.9 mg/mL His<sub>6</sub>-tagged SuiB was mixed 1:1 with precipitant solution to generate a final drop volume of 4  $\mu$ L. Numerous small (< 50  $\mu$ m) star-like clusters of flat rod-shaped crystals were formed within 24 hrs. A seed stock was then produced by combining a single sitting well with 10  $\mu$ L of reservoir solution and 10  $\mu$ L of SuiB at 8 mg/mL, followed by brief vortexing. Seed stock dilutions up to 10<sup>7</sup> were made with the same 1:1 protein/reservoir solution. Crystals were harvested from the 10<sup>6</sup>/10<sup>7</sup> dilutions drops approximately 2 days following seeding. The precipitant solution was 100 mM MES, pH 6.0, 15% (w/v) PEG 3350; cryoprotection was achieved by brief sequential transfer between precipitant solutions with increasing glycerol concentrations of 5%, 10% and 30% (v/v). To obtain the structure with SAM, crystals of SuiB were incubated in precipitant solution containing ~6 mM SAM for 30 min prior to brief sequential transfer between precipitant solutions containing 6 mM SAM and increasing glycerol concentrations of 5%, 10% and 30% (v/v) glycerol. Although spontaneous cleavage of SAM has been observed in the homologous protein StrB and other radical SAM enzymes (1,6), given excess SAM in the soaking condition, a single abortive cleavage event would result in oxidation of the cluster, preventing further activity of the enzyme and yielding intact SAM in the active site.

### 7.8.3 Crystallographic Data Collection and Processing

All data were indexed, integrated and scaled using XDS software followed by merging with AIMLESS (7,8). Model building was completed in COOT (9) and subsequent

refinements/calculations were performed in Phenix (10). Model quality was assessed using Molprobit (11). Data processing and refinement statistics can be found in Table S1. Figures depicting the structure were generated with PyMol. Phasing and Model Building. Single wavelength Fe-anomalous diffraction was collected for a crystal in the absence of substrate at beamline 23-ID-B of the Advanced Photon Source (APS) at Argonne National Laboratory (Chicago, IL) on a MARmosaic 300 CCD detector. The crystal was maintained at 100 K and data were collected using inverse beam ( $\Delta\phi = 1^\circ$ , wedge =  $30^\circ$ ) at the Fe peak ( $\lambda$ , 1.7369 Å). Diffraction approached 2.93 Å with anomalous signal extending to 4.0 Å. Experimental phases were generated with the AutoSol Wizard (12). The hybrid substructure search submodule, HySS, yielded twenty-four heavy atom sites with a figure of merit of 0.39 through 4.0 Å. This is consistent with the presence of three [4Fe-4S] clusters in each of the two asymmetric copies. Solvent flattening was performed with RESOLVE, and the density-modified output map was used to manually generate a basic structure. This model was further augmented using rigid body refinement on a native dataset ( $\lambda$ , 0.6299 Å) from a different crystal collected sequentially ( $\Delta\phi = 0.25^\circ$ ) at the Cornell High Energy Synchrotron Source (CHESS), beamline A1 on a Pilatus 6M (Dectris) detector. The initial model contained two molecules in the asymmetric unit. For structures solved in the absence of peptide substrate (SuiA), higher resolution native datasets ( $\lambda$ , 1.0332 Å) were obtained at 23-ID-D at the Advanced Photon Source (APS). Data were collected sequentially ( $\Delta\phi = 0.1 - 0.2^\circ$ ) at 100 K with the Pilatus 6M (Dectris) detector. Structures of the substrate-free, and SAM-bound enzyme were solved using rigid body refinement of the initial model to 2.5 Å. In each of these structures, there are two molecules in the asymmetric unit; residues that were not modeled due to disorder are

listed in Table S2. Note that the structure of the substrate-free enzyme includes an additional residue at the N-terminus of chain B. The structure of the SAM-bound enzyme includes an intact SAM molecule in each chain. Three [4Fe-4S] clusters were built into each chain of the structures. B-factors near the radical SAM cluster in chain B are consistently higher than the average for the structure. Co-crystallization with SuiA/SAM. A native dataset for reconstituted enzyme cocrystallized with SAM and SuiA was also collected at the APS on the 23-ID-B setup described above. Data were collected sequentially ( $\Delta\phi = 1^\circ$ ) at a wavelength of 1.033 Å. The structure was solved via molecular replacement with the initial enzyme model to yield a 2.1 Å structure with one molecule in the asymmetric unit. The main chain is complete, with methionine and SAM bound in the active site. The substrate peptide, SuiA, is bound, but electron density for residues - 14 and 1 to 8 was insufficient to enable building.

#### 7.8.4 Computational Methods

Four separate simulations were used to investigate the energy landscapes of the modeled core peptide within the determined SuiB crystal structure. In each simulation, we produced an ensemble of spatial starting orientations for the core peptide fragment, whose internal structure was modeled using molecular dynamics constrained by NMR structure-derived constraints (5,7), connected to the leader peptide in the SuiB crystal structure active site. In the first two simulations, referred to as SAMsim\_cycle and SAMsim\_linear, we sampled the core peptide in the presence of an intact SAM found within the solved crystal structure with (cycle) and without (linear) a covalent bond constraint between Lys2 (C $\beta$ ) and Trp6 (C $\zeta$ 2) respectively. The second set of two

simulations, 5ADsim\_cycle and 5ADsim\_linear, were performed in the presence of the cleaved SAM product 5'-deoxyadenosine (5AD) also with and without a Lys-Trp covalent bond present. All 5ADsim simulations were performed with an additional distance constraint added between the core peptide Lys (C $\beta$ ) and the 5'-deoxyadenosine 5'C atom ( $d = 2.8-4.3$  Å). The simulations had the following three steps:

*1) Appending the core peptide with Rosetta Match and Kinematic Loop Closure*

Rosetta Match (13) was used to locate geometrically compatible positions of the core peptide. First, we created a ligand model of the NMR-derived core peptide structure. The ligand model was comprised of the C $\alpha$  atoms of the core peptide structure. Using Rosetta Match and a set of geometric constraints derived from a non-redundant set of high-resolution protein structures (nr database), we located all possible core-peptide C $\alpha$  ligand model placements within the solved crystal structure active site. The ligand model placements positioned the core peptide in a sterically favorable position while maintaining chain connectivity with the leader peptide. To obtain these placements, we applied geometric matching constraints between the C-terminal C $\alpha$  atoms of the leader peptide and N-terminal C $\alpha$  atoms of the core peptide C $\alpha$  ligand model. These geometric constraints were obtained from measuring distance and angle values of contiguous sets of C $\alpha$  atoms within heptapeptide fragments in the nr database. Using these constraints in Rosetta Match, we produced 1296 core peptide placements that did not sterically clash with the SuiB scaffold backbone. For each compatible placement, we converted the C $\alpha$  ligand model to an allatom model and generated a contiguous peptide chain using a generalized kinematic loop closure (genKIC) protocol (14). The genKIC protocol

produced 468 starting structures in which the peptide bond geometries at the connection point were ideal.

### *2) Sampling core peptide conformations within a poly-alanine active site*

In order to enhance the efficiency of conformational space sampling by the core peptide, structures obtained in step 1 were subjected to four cycles of Rosetta FastRelax (15) within a poly-alanine model of the active site using a scoring function that emphasizes the repulsive component of the Lennard Jones potential (16). Residues whose C $\alpha$  atoms were within 8 Å of the core peptide C $\alpha$  atoms were converted to alanine before this step and subsequently returned to their native sidechain conformations after FastRelax. While sampling, we placed NMR derived pseudo-covalent geometry constraints (5,17) between the core peptide residues Lys2 and Trp6, which are involved in the crosslinking reaction, to maintain this covalent linkage. Both the core peptide sidechain and backbone degrees of freedom were sampled. Coordinate constraints were placed on all remaining residues (leader peptide and SuiB-scaffold) to prevent their movement during the simulation. At the end of FastRelax, a final round of rotameric sampling of the core peptide followed by energy minimization (18) was applied with a fixed backbone.

### *3) Refinement of core peptide within the native active site*

We next applied a second round of FastRelax (four cycles), rotameric sampling (four cycles) followed by energy minimization on the conformations generated in step 2. For SAMsim\_cycle and SAMsim\_liner we maintained the crystal structure sidechain conformations in the leaderpeptide and the SuiB-scaffold by placing coordinate

constraints on all crystal-structure residues during the FastRelax and rotameric-sampling stages. Additionally, only the core peptide was allowed to sample both rotameric and backbone degrees of freedom. For 5ADsim\_cycle and 5ADsim\_linear, we performed simulations with and without rotameric sampling of 11 active-site residues (24, 26, 108, 110, 158, 245, 247, 272, 315, 319, and 355). The pseudo-covalent and coordinate constraints from step 2 were maintained in step 3 for SAMsim\_cycle and 5ADsim\_cycle. Additionally, in the 5ADsim\_cycle and 5ADsim\_linear, a second distance constraint of 3.5 Å was placed between the 5'-carbon of 5'-deoxyadenosine and the beta-carbon of Lys2 in the core peptide. All Rosetta scripts and geometric constraint blocks used in the Rosetta simulation are provided below:

1. Matcher constraint block used to locate geometrically compatible conformations of the core peptide with respect to the leader peptide as described in Step 1

```
CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: C1 C2 C3
  TEMPLATE:: ATOM_MAP: 1 residue3: SUI

  TEMPLATE:: ATOM_MAP: 2 atom_name: C CA N
  TEMPLATE:: ATOM_MAP: 2 is_backbone
  TEMPLATE:: ATOM_MAP: 2 residue3: ALA

  CONSTRAINT:: distanceAB: 5.1147 1.036 10.0 0 1
  CONSTRAINT:: angle_A: 111.774 53.71 10.0 360.0 3
  CONSTRAINT:: angle_B: 134.685 25.928 10.0 360.0 3
  CONSTRAINT:: torsion_A: 0.00 180.0 10.0 360.0 12
  CONSTRAINT:: torsion_B: 0.00 180.0 10.0 360.0 12
  CONSTRAINT:: torsion_AB: 0.00 180.0 10.0 360.0 12
  ALGORITHM_INFO:: match
  IGNORE_UPSTREAM_PROTON_CHI
  ALGORITHM_INFO::END
```

```

CST::END

CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: C1 C2 C3
  TEMPLATE:: ATOM_MAP: 1 residue3: SUI

  TEMPLATE:: ATOM_MAP: 2 atom_name: C CA N
  TEMPLATE:: ATOM_MAP: 2 is_backbone
  TEMPLATE:: ATOM_MAP: 2 residue3: SER

  CONSTRAINT:: distanceAB: 6.875 2.875 10.0 0 1
  CONSTRAINT:: angle_A: 105.00 55.00 10.0 360.0 4
  CONSTRAINT:: angle_B: 125.00 45.00 10.0 360.0 4
  CONSTRAINT:: torsion_A: 0.00 180.0 10.0 360.0 12
  CONSTRAINT:: torsion_B: 0.00 180.0 10.0 360.0 12
  CONSTRAINT:: torsion_AB: 0.00 180.0 10.0 360.0 12
  ALGORITHM_INFO:: match
  IGNORE_UPSTREAM_PROTON_CHI
  ALGORITHM_INFO::END
CST::END

CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: C1 C2 C3
  TEMPLATE:: ATOM_MAP: 1 residue3: SUI

  TEMPLATE:: ATOM_MAP: 2 atom_name: C CA N
  TEMPLATE:: ATOM_MAP: 2 is_backbone
  TEMPLATE:: ATOM_MAP: 2 residue3: SER

  CONSTRAINT:: distanceAB: 8.5 4.5 10.0 0 1
  CONSTRAINT:: angle_A: 110.0 70.0 10.0 360.0 5
  CONSTRAINT:: angle_B: 120.0 60.0 10.0 360.0 5
  CONSTRAINT:: torsion_A: 0.00 180.0 10.0 360.0 12
  CONSTRAINT:: torsion_B: 0.00 180.0 10.0 360.0 12
  CONSTRAINT:: torsion_AB: 0.00 180.0 10.0 360.0 12
  ALGORITHM_INFO:: match
  IGNORE_UPSTREAM_PROTON_CHI
  ALGORITHM_INFO::END
CST::END

```

## 2. Generalized kinematic loop closure xml code block used in step 1

```

ROSETTASCRIPTS>
<SCOREFXNS>
  <bb_hbond_tors_fadun_cst weights="empty.wts" symmetric=0>
    <Reweight scoretype=hbond_sr_bb weight=1.17 />
    <Reweight scoretype=hbond_lr_bb weight=1.17 />
    <Reweight scoretype=omega weight=0.5 />
    <Reweight scoretype=rama weight=0.2 />
    <Reweight scoretype=p_aa_pp weight=0.32 />
    <Reweight scoretype=coordinate_constraint weight=10.0 />
    <Reweight scoretype=atom_pair_constraint weight=1.0 />
    <Reweight scoretype=angle_constraint weight=1.0 />
    <Reweight scoretype=dihedral_constraint weight=1.0 />
  </bb_hbond_tors_fadun_cst>
</SCOREFXNS>
<RESIDUE_SELECTORS>
  <Index name=loop resnums=%loop_range% />
  <Not name=not_loop selector=loop />
</RESIDUE_SELECTORS>
<TASKOPERATIONS>
  <InitializeFromCommandline name=init/>
  <IncludeCurrent name=keep_curr/>
  <OperateOnResidueSubset name=nodesrep_notloop selector=loop >
    <PreventRepackingRLT/>
  </OperateOnResidueSubset>
  <OperateOnResidueSubset name=nodes_loop selector=not_loop >
    <RestrictToRepackingRLT/>
  </OperateOnResidueSubset>
</TASKOPERATIONS>
<FILTERS>
  <ContingentFilter name=kicedA_B />
</FILTERS>
<MOVERS>
  <DeclareBond name=bond1 res1=%one% atom1="C" res2=%two% atom2="N"/>
  <GeneralizedKIC name="genkic" closure_attempts=800 stop_if_no_solution=0 stop_when_n_solutions_found=100
  selector="lowest_energy_selector" selector_scorefunction="bb_hbond_tors_fadun_cst" selector_kbt=1.0 contingent_filter="kicedA_B">
    <AddResidue res_index=%one% />
    <AddResidue res_index=%two% />
    <AddResidue res_index=%three% />
    <SetPivots res1=%one% atom1="CA" res2=%two% atom2="CA" res3=%three% atom3="CA" />
    <CloseBond prioratom_res=%one% prioratom="CA" res1=%one% atom1="C" res2=%two% atom2="N"
  followingatom_res=%two% followingatom="CA" bondlength=1.325 angle1=120 angle2=120 randomize_flanking_torsions=true />
    <AddPerturber effect="randomize_alpha_backbone_by_rama">
      <AddResidue index=%one% />
      <AddResidue index=%two% />
      <AddResidue index=%three% />
    </AddPerturber>
    <AddPerturber effect="set_dihedral">
      <AddAtoms res1=%one% atom1="CA" res2=%one% atom2="C" res3=%two% atom3="N" res4=%two%
  atom4="CA"/>
      <AddValue value=180.0/>
    </AddPerturber>
    <AddFilter type="loop_bump_check"/>
  </GeneralizedKIC>
  <AtomCoordinateCstMover name=loopsCST coord_dev=0.2 bounded=true bound_width=0.1 sidechain=true native=false
  task_operations=nodes_loop />
</MOVERS>
<PROTOCOLS>
  <Add mover=bond1/>
  <Add mover=loopsCST/>
  <Add mover=genkic/>
  <Add filter=kicedA_B/>
</PROTOCOLS>
</ROSETTASCRIPTS>

```

3. Enzdes constraint block used to place constraints between Lys2 and Trp6 in the core peptide as well as between core peptide and 5-deoxyadenosine. (Steps 2 and 3)

```

CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: CB CA N
  TEMPLATE:: ATOM_MAP: 1 residue3: LYS

  TEMPLATE:: ATOM_MAP: 2 atom_name: CZ2 CE2 NE1
  TEMPLATE:: ATOM_MAP: 2 is_backbone
  TEMPLATE:: ATOM_MAP: 2 residue3: TRP

  CONSTRAINT:: distanceAB: 1.551 0.030 1000.0 1
  CONSTRAINT:: angle_A: 104.927 5.00 100.0 360.0
  CONSTRAINT:: angle_B: 119.682 5.00 100.0 360.0
  CONSTRAINT:: torsion_A: 179.689 20.00 10.0 360.0
  CONSTRAINT:: torsion_B: 0.813 20.00 10.0 360.0
  CONSTRAINT:: torsion_AB: 70.909 20.00 10.0 360.0
CST::END

CST::BEGIN
  TEMPLATE:: ATOM_MAP: 1 atom_name: CB CA N
  TEMPLATE:: ATOM_MAP: 1 residue3: LYS

  TEMPLATE:: ATOM_MAP: 2 atom_name: C10 C8 N9
  TEMPLATE:: ATOM_MAP: 2 residue3: 5AD

  CONSTRAINT:: distanceAB: 3.500 0.60 100.0 1
CST::END

```

#### 4. XML code block used to run steps 2 and 3 in computational method

```

<ROSETTASCRIPTS>
  <SCOREFXNS>
    <scorefxn1 weights=talaris2013_cst >
      <Reweight scoretype=atom_pair_constraint weight=1.0 />
      <Reweight scoretype=angle_constraint weight=1.0 />
      <Reweight scoretype=dihedral_constraint weight=1.0 />
      <Reweight scoretype=coordinate_constraint weight=1.0 />
      <Reweight scoretype=fa_rep weight=0.1 />
    </scorefxn1>
    <scorefxn2 weights=enxdes_polyA_min.wts />
  </SCOREFXNS>
  <RESIDUE_SELECTORS>
    <Index name=streptide resnums=457-465 />
    <Not name=atoms_with_density selector=streptide />
    <Index name=pocket_to_ala resnums=108,110,112,115,125,126,154,156,158,183,185,208,245,247,249,272,274,277,279,281,282,285,319,348,353,355,357,366,369 />
    <Not name=non_pocket_res selector=pocket_to_ala />
  </RESIDUE_SELECTORS>
  <TASKOPERATIONS>
    <InitializeFromCommandline name=init/>
    <IncludeCurrent name=keep_curr/>
    <OperateOnResidueSubset name=csts_for_non_streptide selector=streptide >
      <PreventRepackingRLT/>
    </OperateOnResidueSubset>
    <OperateOnResidueSubset name=repack_streptide selector=streptide >
      <RestrictToRepackingRLT/>
    </OperateOnResidueSubset>
    <OperateOnResidueSubset name=no_repack selector=atoms_with_density >
      <PreventRepackingRLT/>
    </OperateOnResidueSubset>
    <OperateOnResidueSubset name=pocket_to_ala selector=non_pocket_res >
      <PreventRepackingRLT/>
    </OperateOnResidueSubset>
  </TASKOPERATIONS>
  <FILTERS>
    <PoseInfo name=p_info />

```

```

<ScoreType name=atom_pair_cst scorefn=scorefn1 score_type=atom_pair_constraint threshold=50 />
<ScoreType name=ang_cst scorefn=scorefn1 score_type=angle_constraint threshold=50 />
<ScoreType name=dihedral_cst scorefn=scorefn1 score_type=dihedral_constraint threshold=50 />
<ScoreType name=coord_cst scorefn=scorefn1 score_type=coordinate_constraint threshold=10 />
</FILTERS>

<MOVERS>
<AddOrRemoveMatchCsts name=enzCST cst_instruction="add_new" cstfile="trp_lys.cst" keep_covalent=1 />
ConstraintSetMover name=cstADD add_constraints=true cst_file="all_heavy_atom.cst" />
<AtomCoordinateCstMover name=poseCST coord_dev=0.002 bounded=true bound_width=0.1 sidechain=true native=false task_operations=csts_for_non_streptide />
<AtomCoordinateCstMover name=streptide_bbCST coord_dev=0.6 bounded=true bound_width=0.3 sidechain=false native=false task_operations=repack_streptide />
<MakePolyX name=convert_shell_to_ala keep_gly=1 task_operations=pocket_to_ala />
<SaveAndRetrieveSidechains name=rep_sidechains allsc=1 two_step=1 multi_use=1 jumpid=0 />
<PackRotamersMover name=repack scorefn=scorefn1 task_operations=init,keep_curr,no_repack,repack_streptide />
<MinMover name=min scorefn=scorefn1 chi=1 bb=1 jump=0 cartesian=0 type=lbfgs_armijo_nonmonotone tolerance=0.001 max_iter=200 />

<FastRelax name=fastrelax repeats=4 scorefn=scorefn2 task_operations=keep_curr,init,csts_for_non_streptide >
  <MoveMap>
    <Span begin=457 end=459 chi=1 bb=1 />
    <Span begin=460 end=465 chi=1 bb=0 />
  </MoveMap>
</FastRelax>
<LoopOver name=min_twice mover_name=min iterations=1 drift=true />
<ParsedProtocol name=repack_minimize>
  <Add mover=repack />
  <Add mover=min />
</ParsedProtocol>
<GenericMonteCarlo name=genericMC mover_name=repack_minimize scorefn_name=scorefn1 temperature=0.8 trials=4 />
</MOVERS>
<PROTOCOLS>
  Add mover=cstADD />
  <Add mover=poseCST />
  <Add mover=streptide_bbCST />
  <Add mover=enzCST />

  Add mover=rep_sidechains />
  Add mover=convert_shell_to_ala />
  Add mover=fastrelax />
  Add mover=rep_sidechains />
  Add mover=fastrelax />

  Add filter=p_info />
  Add filter=coord_cst />

  <Add mover=genericMC />
  Add filter=p_info />
  Add filter=atom_pair_cst />
  Add filter=ang_cst />
  Add filter=dihedral_cst />
  Add filter=coord_cst />
</PROTOCOLS>

</ROSETTASCRIPTS>

```

## 7.8.5 SI Tables and Figures

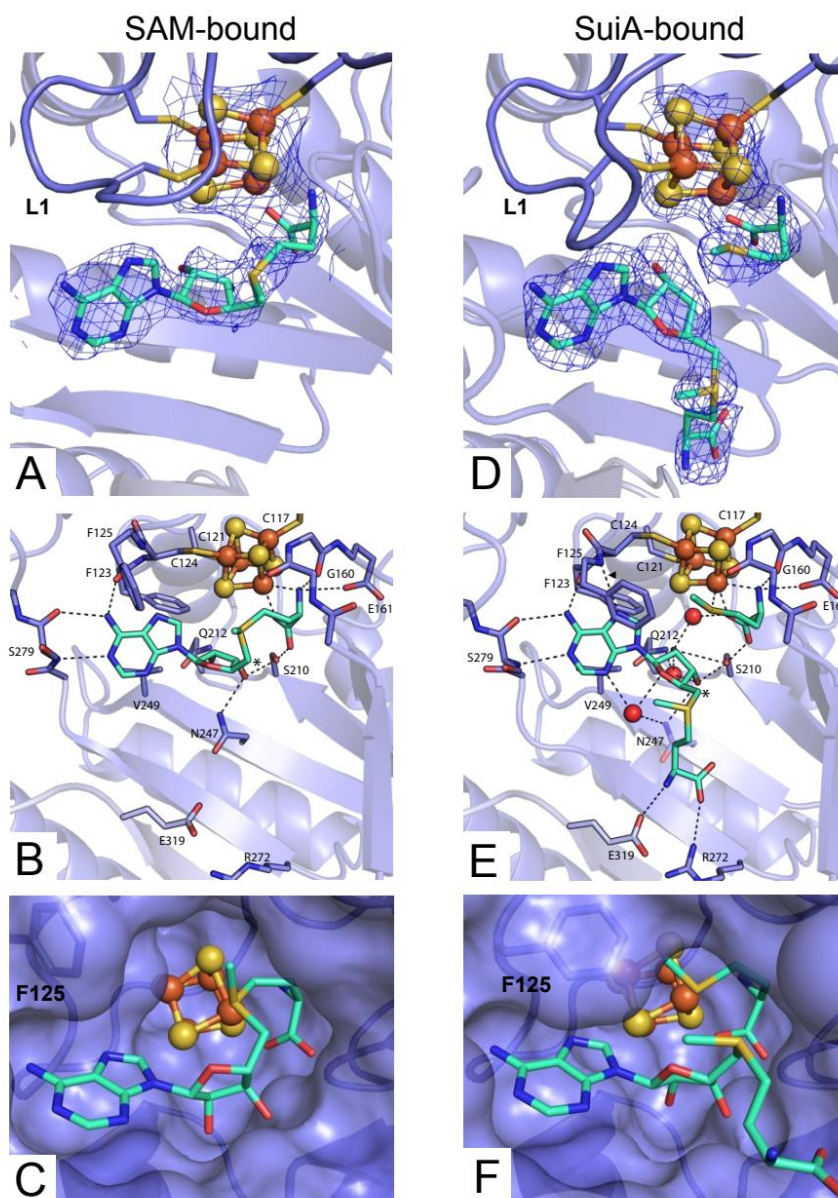
Table 13-7. Table S1. Crystallographic data processing and refinement statistics for SuiB structures.

PDB ID (ligand)	5V1Q	5V1S (SAM)	5V1T (MET/SAM/SuiA)
<b>Data Collection<sup>a</sup></b>			
Space group	P2 <sub>1</sub> 2 <sub>1</sub> 2	P2 <sub>1</sub> 2 <sub>1</sub> 2	P2 <sub>1</sub> 2 <sub>1</sub> 2
Unit cell (Å)	a = 115.23, b = 84.42, c = 110.04 $\alpha = \beta = \gamma = 90$	a = 114.79, b = 85.47, c = 109.86 $\alpha = \beta = \gamma = 90$	a = 69.37, b = 115.04, c = 54.33 $\alpha = \beta = \gamma = 90$
Wavelength (Å)	1.0332	1.0332	1.0332
Resolution range (Å)	29.51 – 2.50 (2.60 – 2.50)	29.47 – 2.49 (2.59 – 2.49)	29.24 – 2.10 (2.16 – 2.10)
Total observations	486901	248251	87295
Total unique observations	37704	38407	25903
I/ $\sigma$ <sub>I</sub>	17.7 (1.9)	17.7 (2.0)	9.3 (1.9)
Completeness (%)	99.3 (94.4)	99.8 (99.0)	99.3 (97.1)
R <sub>merge</sub>	0.096 (1.29)	0.067 (0.839)	0.123 (0.799)
R <sub>pim</sub>	0.028 (0.395)	0.029 (0.373)	0.078 (0.502)
Redundancy	12.9 (11.5)	6.5 (5.9)	3.4 (3.4)
<b>Refinement Statistics</b>			
Resolution range (Å)	29.51 – 2.50	29.47 – 2.49	29.24 – 2.10
Reflections (total)	37648	38359	25862
Reflections (test)	1781	1714	2591
Total atoms refined	6778	6766	3971
Solvent	5	7	283
R <sub>work</sub> (R <sub>free</sub> )	21.66 (25.68)	21.48 (26.97)	18.92 (22.27)
RMSDs			
Bond lengths (Å)/ angles (°)	0.005/0.659	0.008/0.799	0.006/0.942
Ramachandran plot			
Favored/allowed (%)	96.16/3.84	95.24/4.76	96.88/3.12
<b>Mean B values (Å<sup>2</sup>)</b>			
Protein Chains A/B	79.65/79.11	74.36/72.95	31.84/--
[4Fe4S]/SAM/MET/SuiA	66.30/--/--	66.80/71.40/--	25.00/31.70/26.70/36.79
Solvent	71.27	65.40	35.94

<sup>a</sup> Values in parentheses refer to the high-resolution shell.

Table 14-7. Table S2. Missing residues for each structure.

Missing Residues	5V1Q	5V1S	5V1T
Chain A	1, 75-81, 131-133	1, 74-81, 331-335	none
Chain B	none	127-132, 281-286	SuiA(-14), SuiA(1-8)



**Figure 81-7.** Figure S1. SAM binding and cleavage in the SuiB active site.

(A) 2FO-FC composite omit map contoured at  $1.0\ \sigma$  is consistent with an intact SAM bound to the catalytic [4Fe-4S] cluster (Fe – orange, S – yellow). (B) Although the cluster was initially reduced, excess SAM in the absence of reductant yields intact SAM bound in the active site. Hydrogen-bonding network and relevant residues from common radical SAM motifs are labeled. (C/F) Surface renderings of the active site for the SAM-bound and SuiA-bound structures, respectively, depict the formation of a hydrophobic pocket created by changes in loop 1, particularly the perpendicular stacking of F125 with the adenine moiety of SAM. (D) 2FO-FC composite omit map contoured at  $1.0\ \sigma$  is consistent with methionine bound to the catalytic [4Fe-4S] cluster and an intact SAM in the 5'- dA pocket. (E) Post-cleavage

hydrogen-bonding network and motifs that orient methionine and SAM (cyan) in the active site. Note the additional bond formed between F125 and 5'-dA as denoted by the arrowhead. The location of bond cleavage is marked with an asterisk.

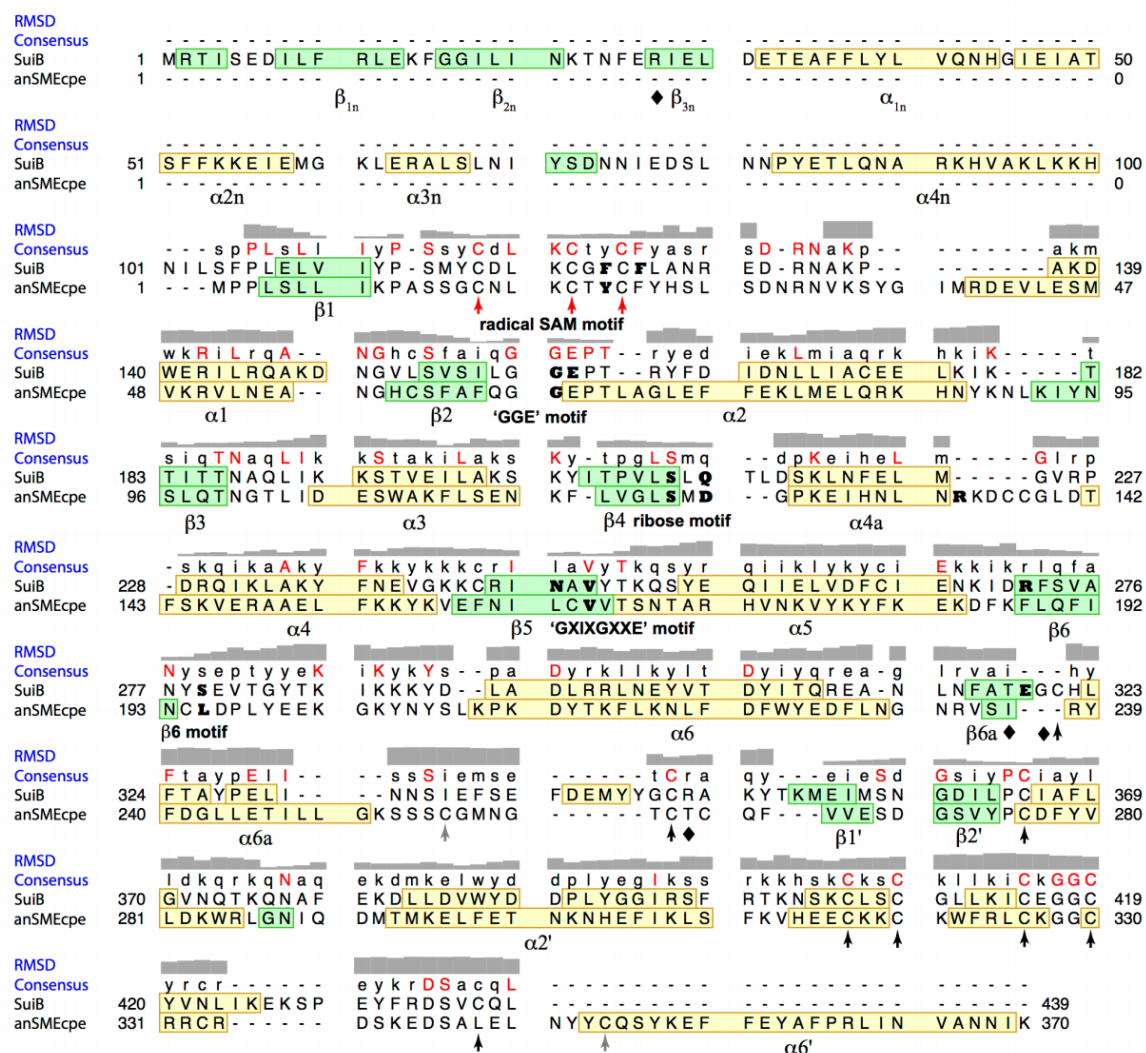
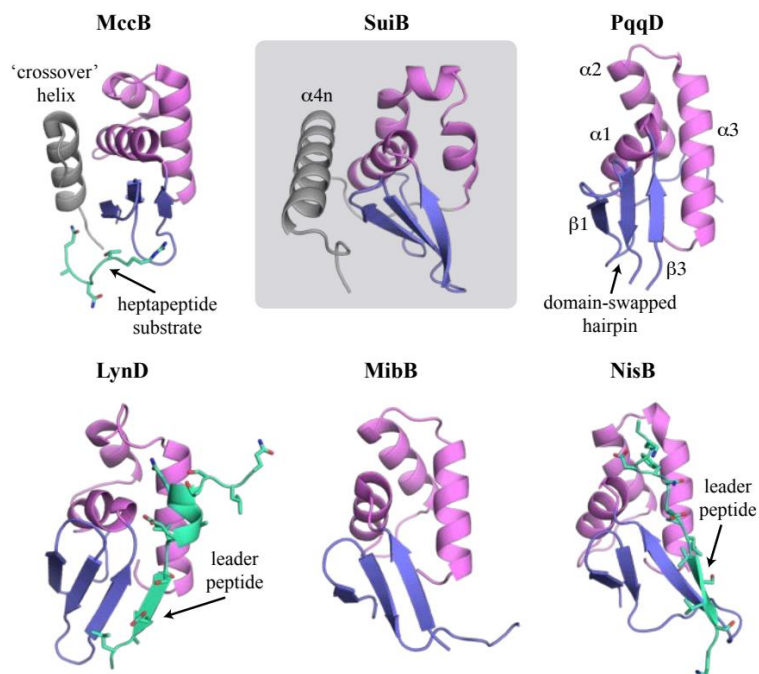


Figure 82-7. Figure S2. Sequence alignment of SuiB with anSMEcpe.

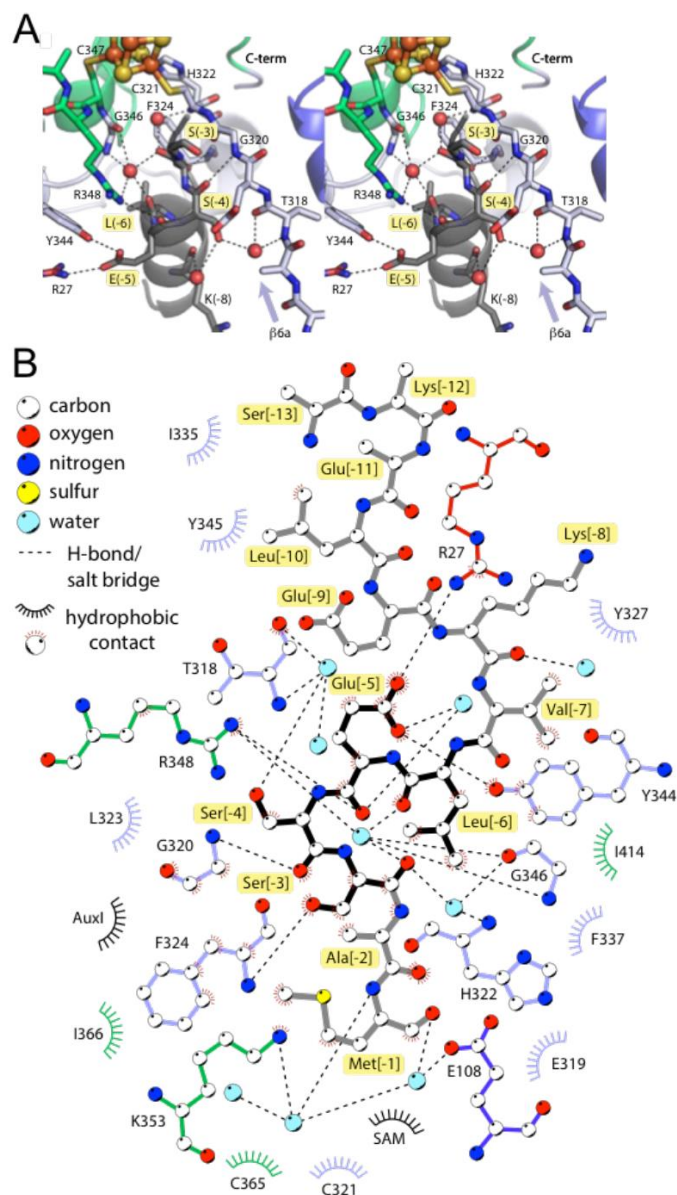
A structure-based sequence alignment was generated using Chimera to yield an overall C $\alpha$  rmsd of 4.82 Å for the aligned 309 residues.  $\beta$ -strands are shown in green and  $\alpha$ -helices in yellow. Primary strands, helices and motifs are labeled below each feature. Secondary structure elements align well for the SAM and SPASM domains (res. 107–310/347–437 in SuiB and res. 3–234/261–348 in anSMEcpe respectively). A histogram, shown in grey, depicts the rmsd by residue, ranging from 0.36 Å to 28.84 Å. Residues that H-bond with SAM or methionine are shown in bold face, whereas black diamonds signify those that directly H-bond with the SuiA leader. Arrows denote Fe/S cluster ligating cysteines. Those corresponding to the SAM cluster are shown in red, the SuiB auxiliary clusters in black, and mismatched cysteines in anSMEcpe in grey.



**Figure 83-7.** Figure S3. Comparison of the RRE domain in SuiB with those previously characterized by X-ray crystallography.

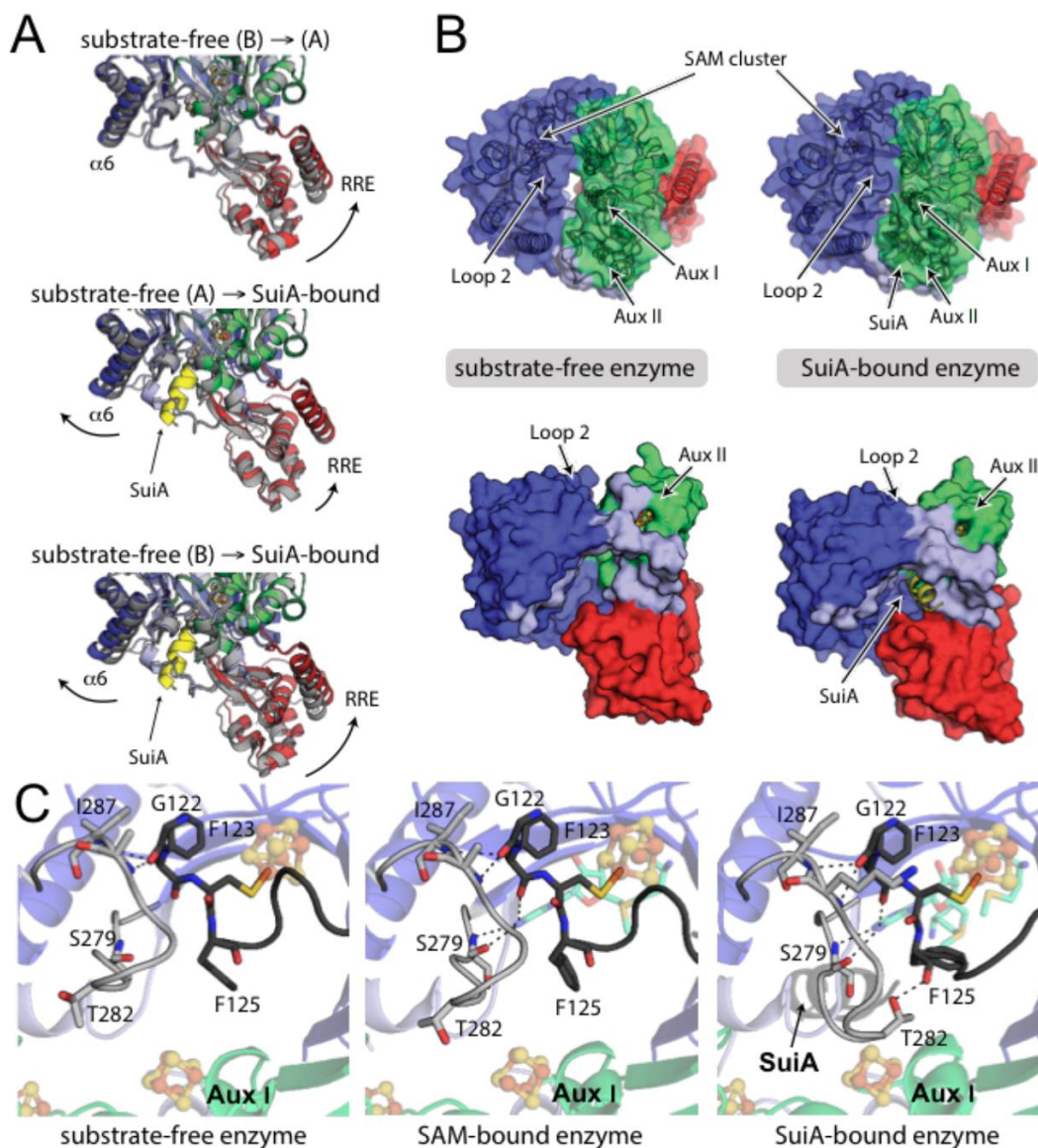
LynD is a fused cyclodehydratase involved in cyanobactin biosynthesis (PDBID: 4V1T); MibB (PDBID: 5EHK) and NisB (PDBID: 4WD9) are lantibiotic dehydratases; MccB (PDBID: 3H9J) is an adenylyase in the microcin C7 biosynthetic pathway; and PqqD (PDBID: 3G2B) is a peptide chaperone involved in the production of PQQ. The characteristic wHTH domain is depicted with purple strands and pink helices; precursor peptides are green. The ancillary helix, corresponding to  $\alpha_{4n}$  in SuiB, is shown in grey and domain-swapped elements of MccB and PqqD are labeled accordingly. The strands and helices in PqqD are labeled for clarity.

Consensus symbols are placed above each residue. An asterisk corresponds to a fully conserved position. A colon (period) indicates strong (weak) agreement between residue properties. Spaces display strong divergence. Strands and helices are shown as cartoons above each feature, with the primary elements labeled. Active site SAM-binding motifs are highlighted in grey, and Fe-ligating cysteines in yellow. Residues that H-bond with SAM or methionine in SuiB are shown in bold face. Hydrogen bonding partners for the leader portion of SuiA are likewise highlighted in blue; all other interacting residues are shown in pink.



**Figure 85-7.** Figure S5. Leader peptide binding site of SuiB.

(A) Stereo view depicting the H-bonding network of SuiA (dark grey) bound in the active site of SuiB. Residues from the bridging domain are shown in light blue, the SPASM in green and the N-terminal RRE domain in red. (B) 2-D protein/peptide interaction map. SuiA labels are highlighted in yellow. Hydrogen bonds are shown for distances less than 3.4 Å. As the electron density terminates immediately after SuiAMet(-1), the direction of the side-chain and corresponding interactions are ambiguous. The displayed orientation was selected for feasibility of peptide continuation into the active site based on the Rosetta simulations.



**Figure 86-7.** Figure S6. Conformational changes in SuiB upon binding of substrate SuiA.

(A) Structural alignment of the substrate-free enzyme chains depict different conformations of the RRE domain and are thus compared independently with the SuiA-bound structure. The structure listed first is shown in grey. SuiA is shown in yellow for clarity. The two chains (A/B) in the asymmetric unit are denoted parenthetically. (B) Surface rendering depicting how the loop movements upon SuiA binding obstruct solvent access to Aux I, Aux II, and the active site. Clusters are shown in ball and stick representation (Fe – orange, S – yellow). (C) Hydrogen-bonding between loops (L1 – black, L2 – light grey) results in coordinated motions upon substrate binding. Upon binding, SAM mediates an additional hydrogen bond between the loops, as shown in Figure S1.

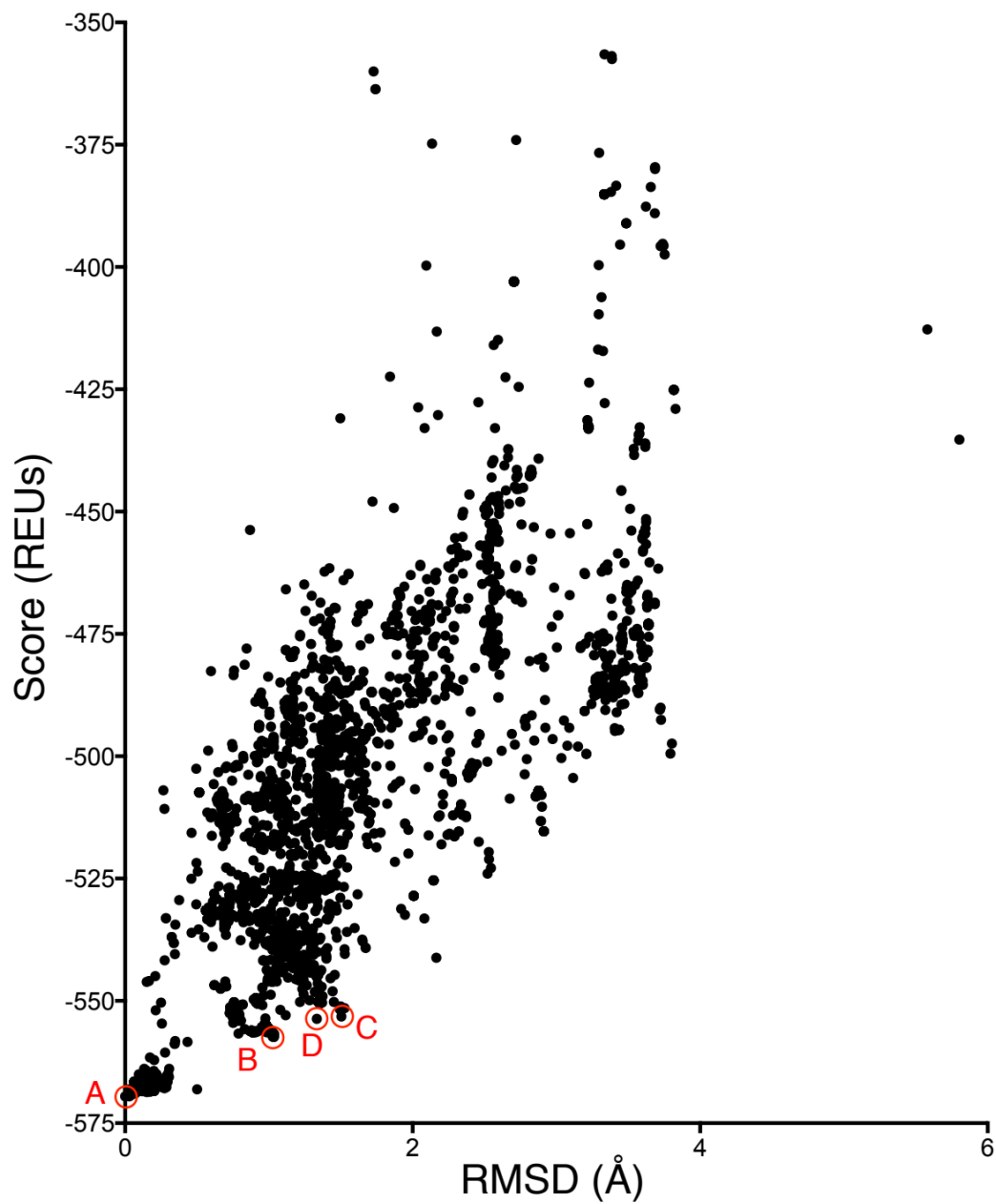
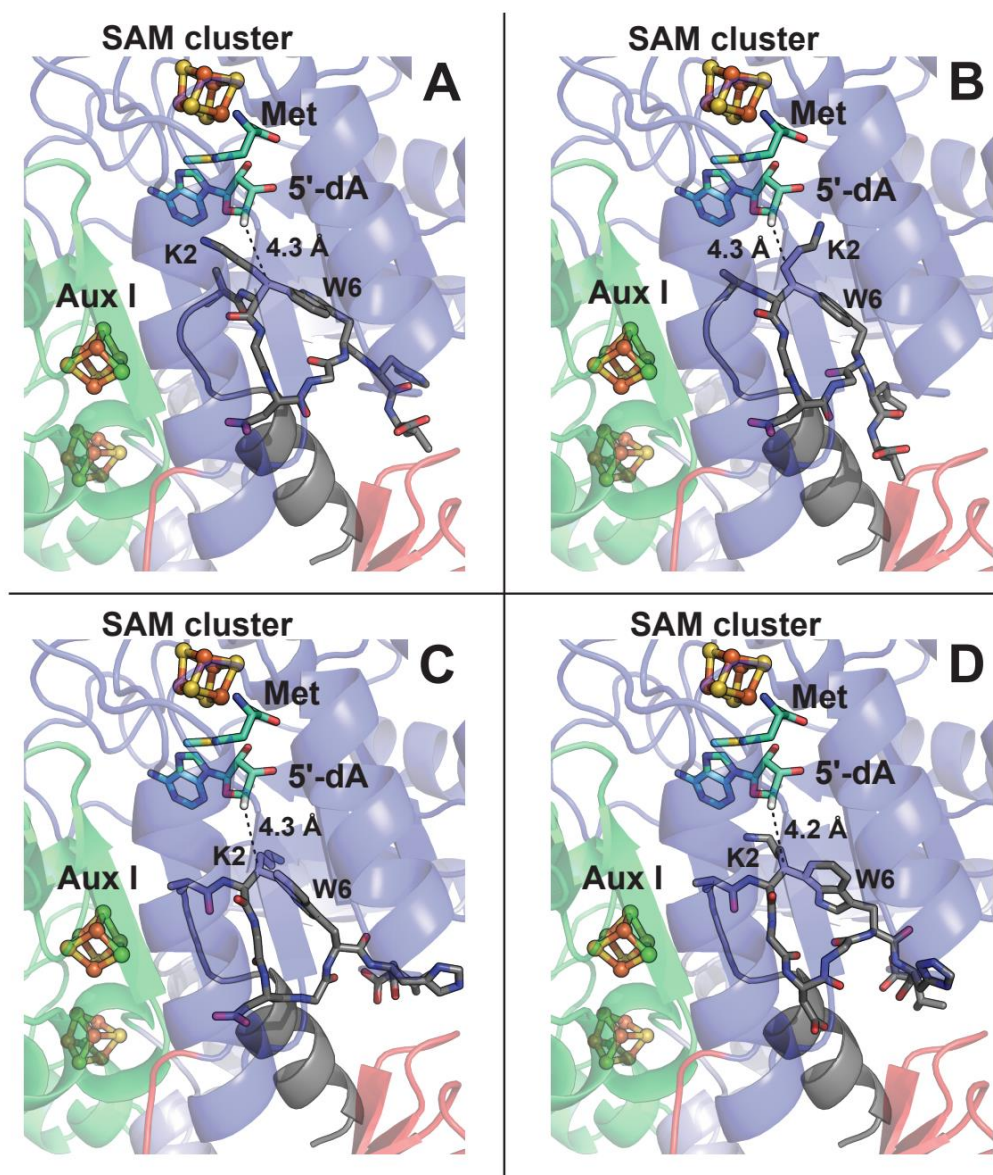


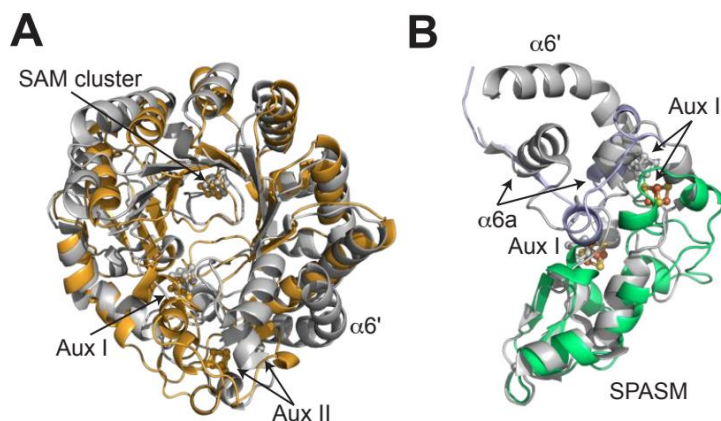
Figure 87-7. Figure S7. Energy landscape of the cyclized SuiA peptide in the SuiB active site.

Score (Rosetta Energy Units) is calculated for the SuiA-SuiB complex for structural models at the end of FastRelax trajectories. RMSD is calculated with respect to the lowest energy models detected in the simulations. Conformations depicted in Fig. S8 are highlighted.



**Figure 88-7.** Figure S8. Rosetta simulations yielded four sets of distinct low energy conformations for the cyclized peptide when SAM was replaced with 5'-dA.

Although very similar, the position of the lysine side-chain and C-terminal residues vary between groups A–C, while in group D the orientation of the indole side-chain is rotated. In all possible conformations, the C-terminus protrudes from the barrel due to space constraints. Hydrogen bonding between SuiA-Asp4 and Arg348 in groups A and B are in agreement with activity assays finding reduced turnover upon an Asp-to-Ala mutation (19).



**Figure 89-7.** Figure S9. Structural comparison with anSMEcpe.

(A) A sequence-independent alignment (RMSD = 3.1Å) of the anSMEcpe (PDB ID: 4K38 - grey) and SuiB (orange) barrels. The RRE domain was omitted for clarity. (B) Expanded view of the SPASM and bridging regions of SuiB (colored) and 4K38 (grey). Rearrangements of the linker region and  $\alpha 6'$  are required due to the binding position of SuiA within the barrel. The radical SAM domain is shown in blue, the linker in light blue, and the SPASM in green. Fe/S clusters are shown in ball and stick representation, where Fe is orange and S is yellow. The peptide substrate, SuiA, is also yellow and depicted in cartoon form.

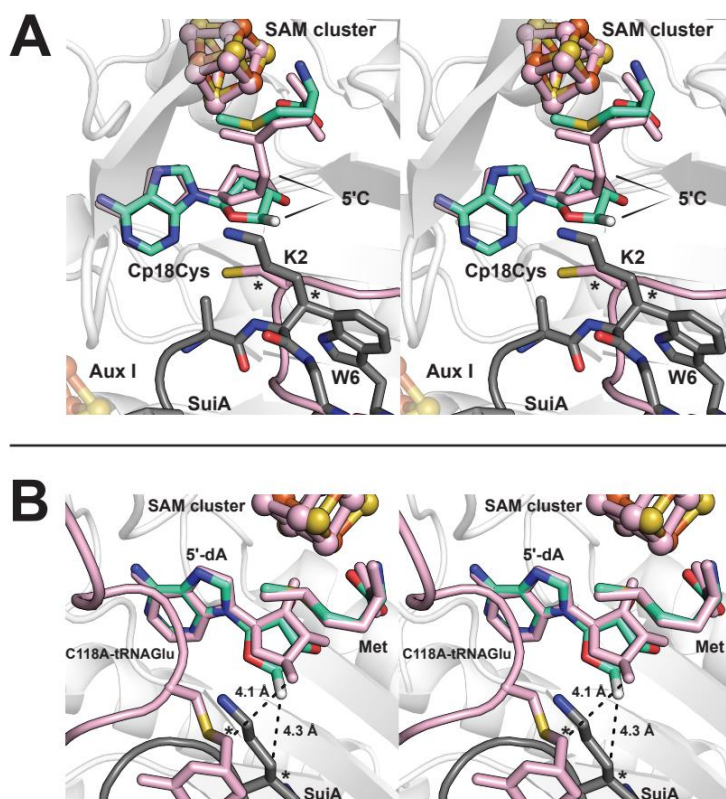


Figure 90-7. Figure S10. The simulated location of the Lys-to-Trp crosslink overlays well with the H-atom abstraction sites of (A) anSMEcpe (20) and (B) RlmN (21).

## **Chapter 8: Stimulus-responsive self-assembly of protein-based fractals by computational design**

### **8.1 Preface**

A version of this chapter has been published in *Nature: Chemistry* and is formatted in the journal style.

### **8.2 Abstract**

Fractal topologies, which are statistically self-similar over multiple length scales, are pervasive in Nature. The recurrence of patterns in fractal-shaped branched objects, such as trees, lungs and sponges, results in a high surface area to volume ratio, which provide key functional advantages including molecular trapping and exchange. Mimicking these topologies in designed protein-based assemblies could provide access to functional biomaterials. Here we describe a computational design approach for the reversible self-assembly of proteins into tunable supramolecular fractal-like topologies in response to phosphorylation. Guided by atomic-resolution models, we develop fusions of Src homology 2 (SH2) domain or a phosphorylatable SH2-binding peptide, respectively, to two symmetric, homo-oligomeric proteins. Mixing the two designed components resulted in a variety of dendritic, hyperbranched and sponge-like topologies that are phosphorylation-dependent and self-similar over three decades ( $\sim 10$  nm– $10$   $\mu$ m) of length scale, in agreement with models from multiscale computational simulations. Designed assemblies perform efficient phosphorylation-dependent capture and release of cargo proteins.

### 8.3 Main

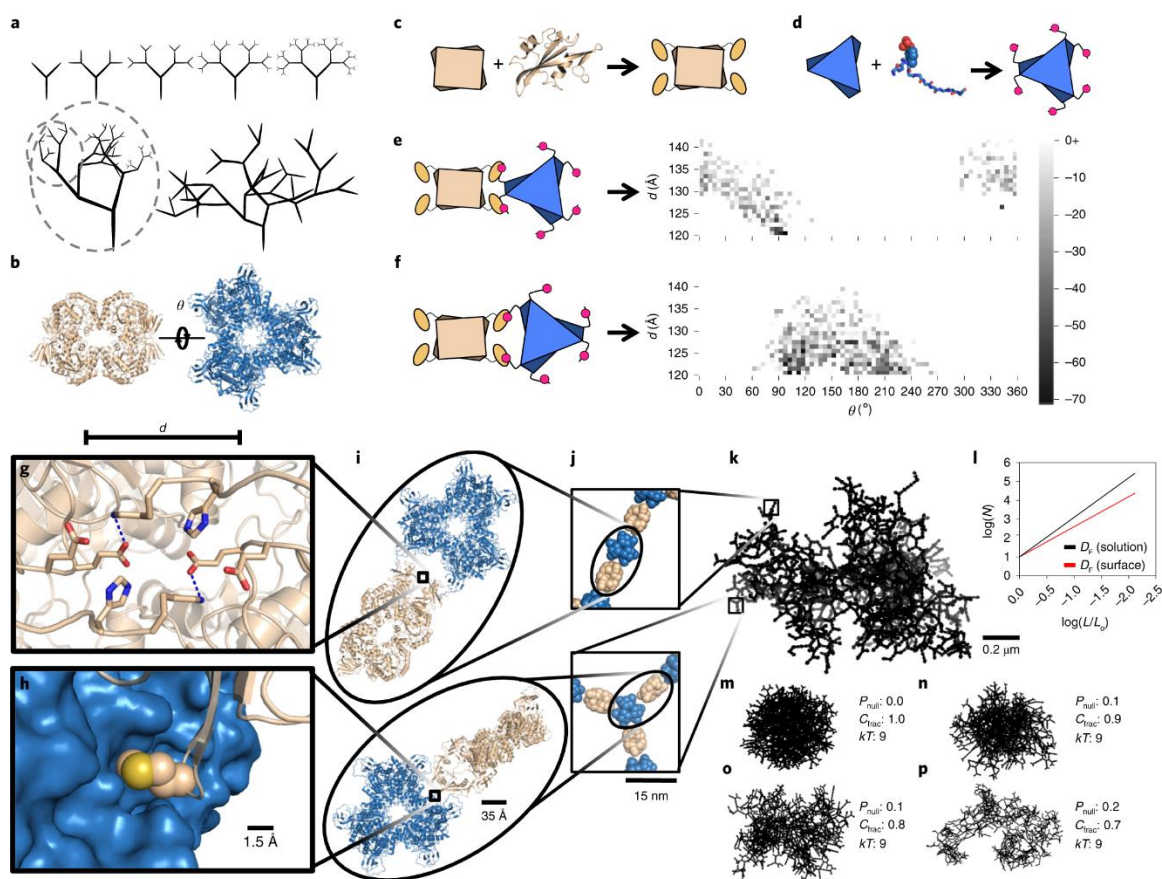
Fractional-dimensional (fractal) geometry—a property of shapes that are invariant or nearly invariant to scale magnification or contraction across many length scales—is a common feature of many natural objects<sup>1,2</sup>. Fractal forms are ubiquitous in geology (for example, in the architecture of mountain ranges), as well as in coastlines, snow formations and in physiology (for example, in neuronal and capillary networks and nasal membranes, where highly efficient molecular exchange occurs due to a fractal-induced high surface area to volume ratio<sup>3</sup>). Fabrication of fractal-like nanomaterials affords high physical connectivity within patterned objects<sup>4</sup>, ultrasensitive detection of target binding moieties by patterned nanosensors<sup>5</sup>, and rapid exchange and dispersal of energy and matter<sup>6</sup>. An intimate link between structural fractal properties of designed, nanotextured materials and functional advantages (for example, detection sensitivity) has been demonstrated<sup>5</sup>, and synthetic fractal materials are finding applications in sensing, molecular electronics, high-performance filtration, sunlight collection, surface charge storage and catalysis, among myriad other uses<sup>7,8</sup>. Many fractal fabrication efforts have relied on top-down patterning of surfaces<sup>9</sup>. The bottom-up design of supramolecular fractal topologies—both deterministic (for example, Sierpinski's triangles)<sup>10,11</sup> and stochastic (for example, arborols)<sup>12,13</sup> fractals—has been performed with small-molecule building blocks such as inorganic metal–ligand complexes or synthetic dendritic polymers utilizing coordinate or covalent bonds, respectively. Self-similar quasi-fractal shapes built with DNA origami have been reported<sup>14,15,16</sup>; however, fractal topologies have not been designed with proteins, which possess a wide range of functionalities and biocompatibility, and whose properties are dynamically controllable by reversible post-

translational modifications<sup>17</sup>. Although fractal-like topologies have been detected as intermediates in the formation of natural protein-based biomaterials such as biosilica and silk<sup>18,19</sup>, and have also been observed in peptide assemblies<sup>20,21,22</sup>, their tunable construction by utilizing reversible non-covalent interactions between protein building blocks under mild conditions remains a fundamental design challenge.

Self-assembly of engineered proteins<sup>23</sup> provides a general framework for the controllable and bottom-up fabrication of novel biomaterials with chosen supramolecular topologies, but these approaches have thus far only been applied to the design of integer (two or three)-dimensional ordered patterns such as layers, lattices and polyhedra<sup>24,25,26,27,28,29,30</sup>. Although external triggers such as metal ions and redox conditions have been used to trigger synthetic protein and peptide assemblies<sup>20,21,31,32,33,34</sup>, phosphorylation—a common biological stimulus used for dynamic control over protein function—has yet to be utilized for controlling protein assembly formation.

Among stochastic fractals, an arboreal (tree-like) shape is an elementary topology that can be generated using stochastic branching algorithms, such as L-systems<sup>35,36</sup>, in which the probability of branching, length and number of branches, and branching angle ranges at each iteration determine the emergent topology (Fig. 1a). Theoretical and simulation studies on the self-assembly of 'patchy' colloidal particles<sup>37,38</sup> have shown that a variety of topologies, including fractal-like topologies<sup>39,40,41</sup>, can result from stochastic self-assembly processes involving strong, anisotropic short-range forces<sup>42,43,44,45,46</sup>. Under conditions where the inter-molecular interaction energy is much larger (more negative) than the thermal energy, emergent large-scale aggregates are expected to be out-of-

equilibrium kinetically trapped states rather than (usually crystalline) globally stable thermodynamic minima. As the reorganization of aggregate morphologies, once formed, is expected to be unfavourable, we reasoned that these kinetic traps can be utilized to produce a tunable and responsive structural (and thus functional) diversity of self-assembled protein-based systems<sup>47</sup>.



**Figure 91-8.** Fig. 1: Multiscale computational design approach for fractal assembly design.

**a**, Cartoon representations of an ordered self-similar scaling fractal, an unordered self-similar scaling fractal—note concentric circles are self-similar at different scales—and an unordered statistically self-similar fractal. **b**, A two-component library of AtzC (tan) and AtzA (blue) positions was generated by varying the rigid body degrees of freedom along paired  $C_2$  symmetry axes. **c, d**, Design and modelling of assembly at the molecular scale was performed by fusing an SH2 binding domain and its corresponding phosphorylatable peptide to AtzC (c) and AtzA (d), respectively. The linker between the SH2 domain and AtzC was designed to ensure symmetric binding between the hexamer and tetramer

leading to propagation. **e,f**, Flexibility analysis was performed by evaluation of the Rosetta energy landscape of symmetrical connections and the probability of observing different connection distances and angles was calculated using the Boltzmann distribution for two binding modes: vertex (**e**) and edge (**f**). Boltzmann-weighted connection probabilities were utilized in a stochastic chain-growth program with a coarse-grained protein model to generate emergent structures. **g–k**, Atomic interactions that stabilized novel interfaces formed from physically connected components (**g,h**) dictate the rotation along the *C*-symmetric axis between components (**i,j**), which ultimately produce combinations of orientations that lead to fractal-like topologies (**k**) on the micrometre scale. **l**, Representation of expected fractal dimension (slope) for fractals analysed in solution and on surfaces. The fractal dimension is calculated using a box-counting approach in which the image is divided into boxes of increasing length.  $N$  is the number of boxes that contain pixels at each box length,  $L$ , and  $L_o$  is the largest possible box (starting box) that fits the image. **m–p**, Examples of fractal simulation output across varying termination probabilities ( $P_{\text{term}}$ ) and fraction of components ( $C_{\text{frac}}$ ) at fixed  $kT$ .

## 8.4 Results

To implement a general approach for tunably designing arboreal fractal morphologies using triggerable self-assembly of protein building blocks, we envisioned the need for (1) a set of branching components whose binding to each other would lead to propagation of the assembly (Fig. [1a](#)), (2) a modular system for connecting, with high affinity, these components reversibly in response to a chosen chemical trigger and (3) degeneracy of protein–protein binding modes (geometries), such that stochastic, but anisotropic, directional propagation of multiple branching geometries leads to emergent fractal-like supramolecular topologies (Supplementary Fig. [1](#)). We chose (1) the oligomeric enzymes AtzA (hexameric) and AtzC (tetrameric) of the atrazine biodegradation pathway<sup>[48](#)</sup> featuring dihedral ( $D_3$  and  $D_2$ , respectively) symmetry (Fig. [1b](#)), (2) a phosphorylatable peptide (pY) tag with its corresponding engineered high-affinity 'superbinder' Src homology 2 (SH2) domain<sup>[49](#)</sup> and (3) linker segments that can stabilize multiple binding orientations, respectively, as design elements encoding these properties (Fig. [1b–d](#)). We

have previously utilized a similar binding domain–peptide fusion strategy to design non-propagating multicomponent enzyme complexes<sup>50</sup>.

#### 8.4.1 Computational design and multiscale modelling of assembly formation

The sequences of the designed protein components were obtained using a procedure implemented in the Rosetta macromolecular modelling program<sup>51</sup>, aimed at making a maximum of three divalent connections between each AtzA and AtzC mediated by SH2 domain-phosphopeptide binding (Fig. [1c,d](#)). Divalent connections between components were sought to enable avidity, leading to strong, directional, short-range interactions ('aeolotropic interactions'<sup>44</sup>) that would promote fractal growth (Fig. [1e,f](#)). We also reasoned that geometric degeneracy in the form of multiple propagatable (but still anisotropic) binding modes would favour fractal structures (Supplementary Fig. [1](#)). In the first step of the design procedure, one of the  $C_2$  axes of the crystallographic structures of the two components were aligned (Fig. [1b](#)). Two alignments (Fig. [1e,f](#)), obtained by rotating AtzA (hexamer) by  $180^\circ$  about its  $C_3$  axis, were considered, and the remaining two symmetry-compatible degrees of freedom for placement—the inter-component centre-of-mass distance  $d$  and rotation angle  $\theta$  about the aligned axis of symmetry—were sampled (Fig. [1b,e,f](#)). For every value of  $d$  we sampled several discrete values of  $\theta$  that, if uniformly adopted, were predicted to lead to an infinitely propagatable integer-dimensional lattice (Supplementary Fig. [1](#)). The resulting propagatable placements were evaluated using RosettaMatch<sup>31</sup> for geometrically feasible fusion to the SH2 domain and phosphopeptide with the C-terminal AtzC and N-terminal of AtzA, respectively (Supplementary Fig. [2](#)). Loop closure of successful SH2 domain and phosphopeptide

placements was performed using Rosetta kinematic loop closure (Supplementary Fig. 2). Next, optimization of the new intra- and inter-component interfaces was performed using RosettaDesign (Supplementary Fig. 3). Five AtzA–AtzC fusion protein pairs were chosen for experimental characterization based on removal of steric clashes (as reflected by the calculated Rosetta energy, Supplementary Table 1), tight interface packing between the SH2 domain and AtzC, and visual examination of design models (Supplementary Fig. 3). We found that short, flexible linker sequences (for example, Gly-Gly-Ser) between the SH2 domain and AtzC led to the most efficient interface packing in designs, while still potentially allowing multiple binding modes: several mutations were common among design models obtained at different (single) values of  $(d, \theta)$ , suggesting geometric degeneracy in binding by each variant (Supplementary Table 1) would be feasible. Indeed, several other values from the propagatable angle set are energetically feasible for each designed AtzC-SH2 variant (Supplementary Table 1).

To fully evaluate the predicted geometric degeneracy and anisotropy of binding in designed inter-component interactions, the conformational landscape over all  $(d, \theta)$  pairs (Fig. 1e,f) was constructed using Rosetta SymmetricFastRelax simulations for a designed hexamer–tetramer complex, and the calculated energies were Boltzmann-weighted (using a simulation temperature parameter,  $T$ ) to obtain a probability distribution  $P(d, \theta)$  for the branching geometry. This distribution, in turn, was used as input for a coarse-grained stochastic chain-growth tree generation algorithm for predicting ensembles of emergent topologies on the micrometre length scale. Similar hierarchical approaches have previously been developed for modelling protein crystallization<sup>52</sup> and colloidal particle<sup>43</sup> and protein self-assembly<sup>45</sup>. In our approach, preferred inter-component interaction

modes at the sub-nanometre scale (Fig. [1g,h](#)) guide the emergence of higher-order structures on the nanometre (Fig. [1i,j](#)) and micrometre length scales (Fig. [1k](#)). For comparison with experiments, ~100s of emergent structures in the resulting ensemble were analysed to determine the fractal dimension ( $D_F$ ) using the box counting image processing technique (Fig. [1l](#)). The fractal dimension of an object is a measure of how its mass or shape scales as a function of length scale (Supplementary Fig. [4](#)): an object is considered fractal if this scaling exponent is non-integer and typically less than the Euclidean dimension in which the object is placed. For example, the  $D_F$  of vasculature patterns on the two-dimensional (2D) surface of the human retina<sup>[53](#)</sup> is 1.7, and a diffusion-limited aggregation cluster in 3D space has a  $D_F$  of 2.3 (ref. [54](#)). In our simulations, a variety of assembly sizes and fractal dimensions,  $D_F$ , could be obtained by varying three parameters: (1) the fraction of growth sites selected at each growing layer allowed to continue propagation ( $c_{\text{frac}}$ ), which reflects the stoichiometry of the two components, (2) the probability of termination at any chosen propagatable branching point ( $P_{\text{term}}$ ), which reflects the affinity of interactions (the lower the affinity, the higher the  $P_{\text{term}}$ ) and (3) the Boltzmann factor ( $k_B T$ ), which determines the sampling of inter-component conformational diversity calculated from Rosetta simulations (Fig. [1m-p](#) and Supplementary Figs. [5](#) and [6](#)).

#### 8.4.2 Experimental characterization of designed assemblies

Genes encoding the designed AtzA and AtzC variants and the corresponding fusions of wild-type domains were constructed and cloned into an *Escherichia coli* BL21(DE3) strain harbouring a second plasmid for the inducible expression of GroEL/ES chaperones

to aid protein yields. Purified AtzA designs were each phosphorylated using Src kinase, and the presence of phosphotyrosine was confirmed by enzyme-linked immunosorbent assay (ELISA; Supplementary Fig. [7](#)); binding and assembly formation with purified AtzC-SH2 domain fusions was assessed using biolayer interferometry and dynamic light scattering (DLS), respectively. Phosphorylation, binding and complete conversion of monomers into 1–10  $\mu\text{m}$  particles upon mixing was best detected with the proteins pY-AtzAM1 and AtzCM1 (Supplementary Figs. [8–10](#)). Either phosphorylation levels were lower or inter-component binding was weaker (Supplementary Fig. [9](#)) with other designs, so we chose the pY-AtzAM1:AtzCM1 design pair for further characterization of assembly–disassembly processes (Fig. [2a](#)). Apart from fusion of pY-tag and SH2 domain, these proteins feature two and three substitutions compared to their wild-type parent, respectively (Supplementary Table [1](#) and Supplementary Figs. [11](#) and [12](#)).

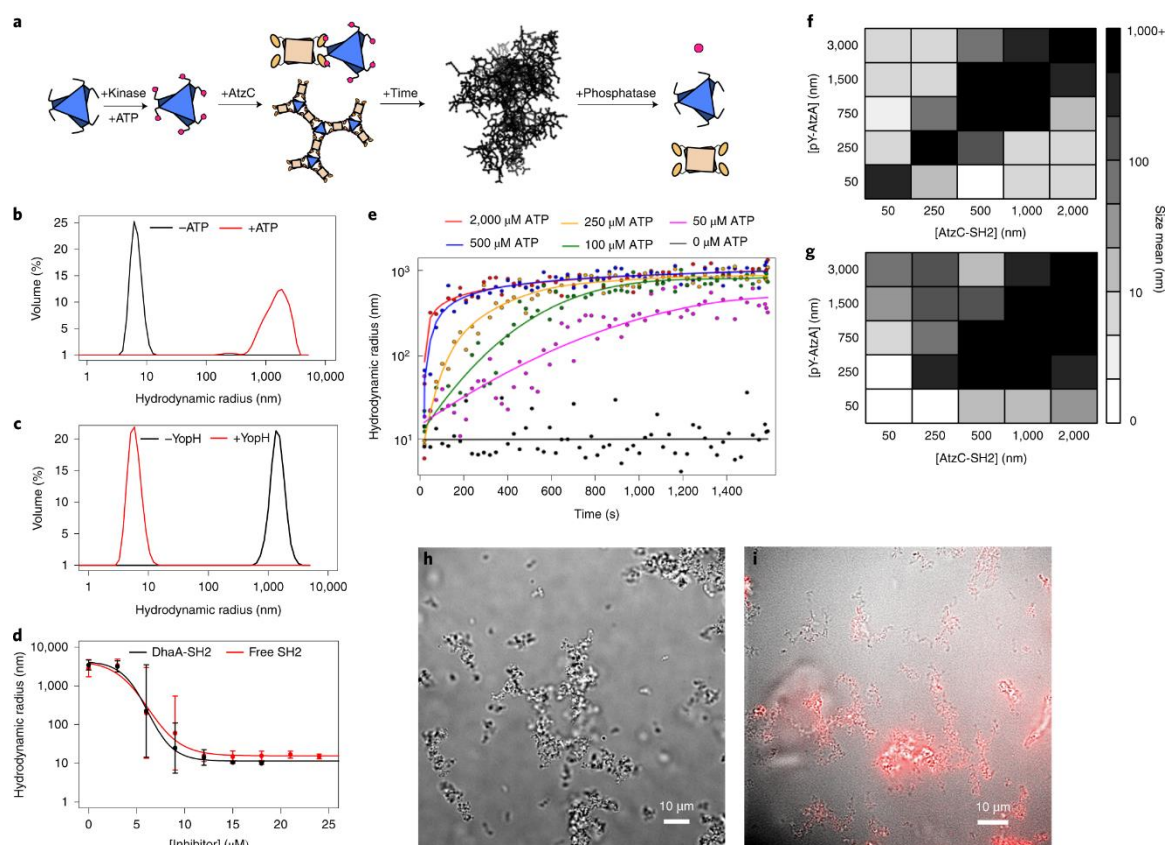


Figure 92-8. Fig. 2: Assembly formation, dissolution and inhibition in vitro.

**a**, Using Src kinase, the AtzAM1 can be phosphorylated (pY-AtzAM1) and incubated with AtzCM1-SH2 to form an assembly. The phosphatase (YOP) enzyme can be used for disassembly. **b,c**, Assemblies were expected to form (**b**) and dissolve (**c**), respectively, as confirmed by DLS measurements. **d**, Incubation of assembling components with various concentrations of free SH2 domain and a different (monovalent) SH2 fusion protein led to robust inhibition. **e**, ATP concentration was shown to control the rate of assembly formation (highest concentration of ATP to lowest, starting from top to bottom at time 0) **f,g**, Assembly formation is highly sensitive to the stoichiometry of the components. Varying the stoichiometry (**f,g**) and the use of a weaker-binding SH2-peptide interaction (**f**) leads to a perturbation of the assembly formation zone compared to the 'superbinder' SH2 (**g**). **h**, Fractal-like structure observed by light microscopy. **i**, Fluorescence microscopy image of assembly formed by Alexa Fluor 647-labelled AtzCM1-SH2 and pY-AtzAM1.

Assembly formation by a mixture of the two components and Src kinase enzyme was adenosine triphosphate (ATP)-dependent (Fig. 2b) and was accompanied by the visible and spectrophotometrically measurable (Supplementary Fig. 13) appearance of turbidity, which could be reversed by adding a phosphatase (YopH) enzyme. The resulting

distribution of particle sizes was detected by measuring the hydrodynamic radii using DLS (Fig. [2c](#)). On completion of assembly formation, the apparent size of the particles as measured by DLS was between 1 and 10  $\mu\text{m}$ ; however, this range represents the upper limit of measurement for the instrument, and actual particle sizes are expected to be larger. Addition of monovalent competitive inhibitors, that is, isolated SH2 domain or SH2 domain fused to an unrelated monovalent protein (SH2-DhaA), inhibited assembly formation in a concentration-dependent manner, demonstrating that the SH2-pYtag binding interaction underlies assembly formation. The apparent half-maximum inhibitory concentration for the observed inhibition was  $\sim 2 \times [\text{AtzA-pY}]$  (measured as monomers) at two different concentrations of the components (Fig. [2d](#) and Supplementary Figs. [14–16](#)), and in each case  $\sim 3 \times [\text{AtzA-pY}]$  was required for complete inhibition. According to our design model, each pY-AtzA (hexamer) makes at least two and at most three divalent connections for assembly propagation (Fig. [1e,f](#)); thus, the observed inhibition stoichiometries are consistent with the existence of the designed divalent connections between AtzA-pY and AtzC-SH2 in the assemblies.

As the phosphorylation reaction requires ATP, assembly formation rates could be controlled by varying the concentration of added ATP. For  $[\text{AtzA-pY}]$  and  $[\text{AtzC-SH2}]$  of 3  $\mu\text{M}$  and 2  $\mu\text{M}$ , respectively,  $[\text{ATP}] > 250 \mu\text{M}$  led to complete conversion of monomers to assemblies within 5 min, whereas significantly slower rates of conversion were observed with lower  $[\text{ATP}]$  (Fig. [2e](#), Supplementary Fig. [17](#) and Supplementary Table [2](#)). Visualization of assemblies using optical and fluorescence microscopy (with Alexa-647-labelled AtzC-SH2) revealed the existence of large ( $>10 \mu\text{m}$ ) dendritic structures (Fig. [2f,g](#)), whose formation could be observed in real time by adding kinase

and ATP to a mixture of the two component proteins (Supplementary Video [1](#) and Supplementary Fig. [18](#)).

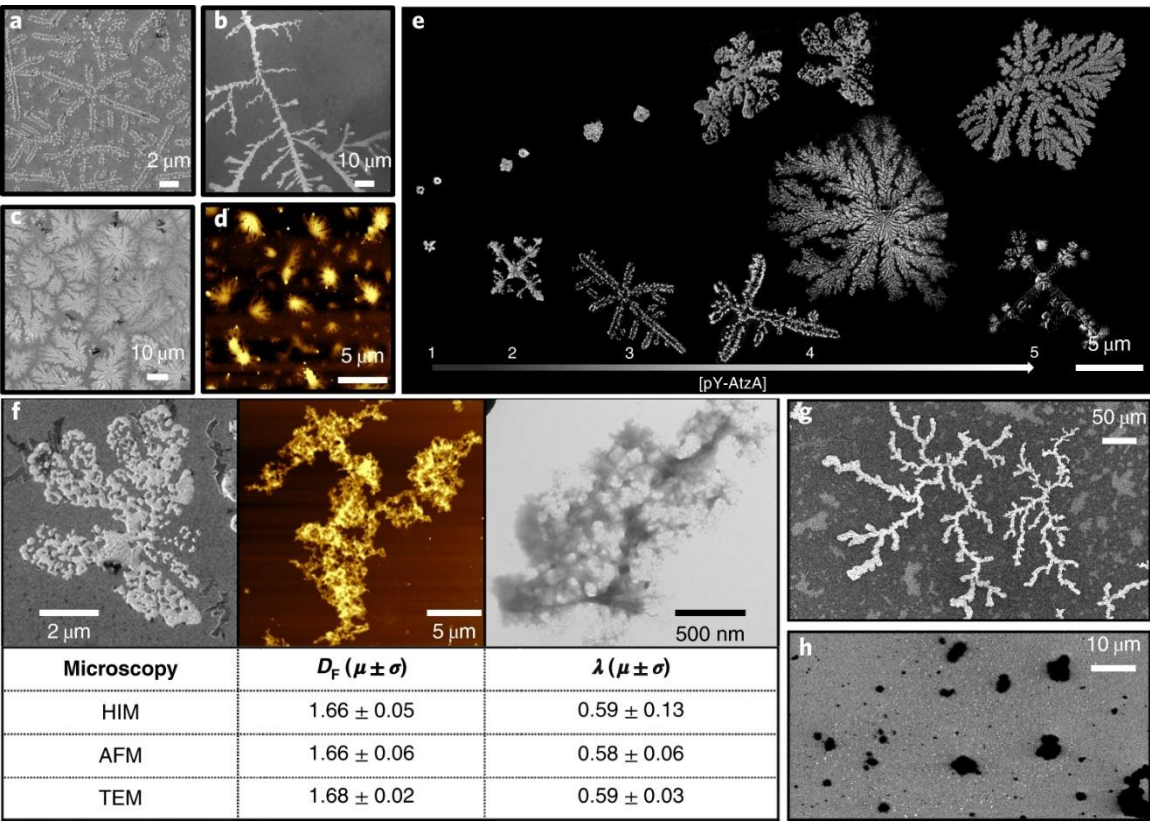
Apparent hydrodynamic radius (Fig. [2f,g](#)) and polydispersity measured with DLS (Supplementary Figs. [19](#) and [20](#)) could be controlled by varying the relative stoichiometry of the two components and by using a weaker binding affinity variant of the SH2 domain fused to AtzC. A comparison of assembly formation trends for the lower- (Fig. [2f](#)) and higher-affinity (Fig. [2g](#)) SH2-domain-containing constructs shows that robust assembly formation is observed at nearly equal concentrations of the two components. Assemblies can be formed at concentrations as low as 50 nM (dissociation constants  $K_D$  for the weaker and tighter interactions were measured as ~40 and ~7 nM, respectively; Supplementary Fig. [10](#)), whereas when one component is present in excess, assembly formation is inhibited, as expected from our branch propagation design model (Fig. [1](#)). Assembly formation by non-stoichiometric concentration combinations with the higher-affinity SH2 domain variant (Fig. [2f,g](#)) indicates that the inhibition caused by an excess of the binding partner is dynamic. Inhibition of assembly formation due to stoichiometric excess can also be overcome in an affinity-dependent manner: the zone of stoichiometries where assembly formation occurs is larger for the higher-affinity SH2 domain variant (Fig. [2g](#)) compared to the lower-affinity variant (Fig. [2f](#)). These results highlight the importance of high affinity in stabilizing the designed kinetically trapped aggregate state: under conditions of uneven stoichiometry (for example, 250 nM AtzA-pY; 1,000 nM AtzC-SH2) and in the absence of kinetic traps, all AtzA components should be bound by an excess of AtzC-SH2 domains, and no assemblies should result (expected particle size is <50 nm). This behaviour is observed for the weaker-affinity

SH2 variant at this stoichiometry (Fig. [2f](#)). In stark contrast, for the high-affinity SH2 variant (Fig. [2g](#)), we observe micrometre-sized assemblies indicating the presence of aeolotropic kinetic trapping<sup>44</sup> and network formation by clusters of tightly bound AtzA-pY-AtzC-SH2 assemblies (Supplementary Video [1](#)).

#### 8.4.3 Structural characterization of surface-adsorbed assemblies

We next investigated if the dynamic and dendritic structures observed in solution by optical and fluorescence microscopy (Fig. [2h,i](#)) could form fractals on solid surfaces, and if the topology of the surface-adsorbed assemblies could be controlled by varying the component stoichiometry. Due to the substantial increase of surface area derived from fractal patterns, surface-adsorbed fractals at the nanometre–micrometre scale are attractive design targets for applications in many fields, such as catalysis, fractal electronics and the creation of nanopatterned sensors<sup>4,5</sup>. Assemblies with a chosen stoichiometry of components were generated in buffer, dropped on the surface of a silicon (or mica) chip, and the solvent was evaporated at room temperature (298 K) under a dry air atmosphere. Visualization of these coated surfaces using helium ion and atomic force microscopy (AFM) reveals striking, intricately textured patterns that coat areas of up to 100  $\mu\text{m}^2$ . Various morphologies on the micrometre scale—including rod-like, tree-like, fern-like and petal-like—were observed (Fig. [3a–e](#)); image analysis revealed fractal dimensions between 1.4 and 1.5 (Fig. [3a,b](#)) and to the more diffusion limited aggregation (DLA)-like 1.78 (Fig. [3c,d](#) and Supplementary Figs. [21](#) and [22](#)). Assembly sizes and fractal dimensions could be tuned by varying the stoichiometry of the components (Fig. [3f](#)), although some heterogeneity in morphologies was present in each sample. At

1:1 stoichiometry of the two components, DLA-like topologies with ~10 μm size were observed, whereas more dendritic assemblies were observed when unequal stoichiometry samples were used (Fig. 3f). Similarly, smaller assembly sizes resulted when the concentration of one component became limiting.



**Figure 93-8.** Fig. 3: Assembly formation and characterization with helium ion microscopy, AFM and transmission electron microscopy. All methods reveal fractal-like topologies on a surface.

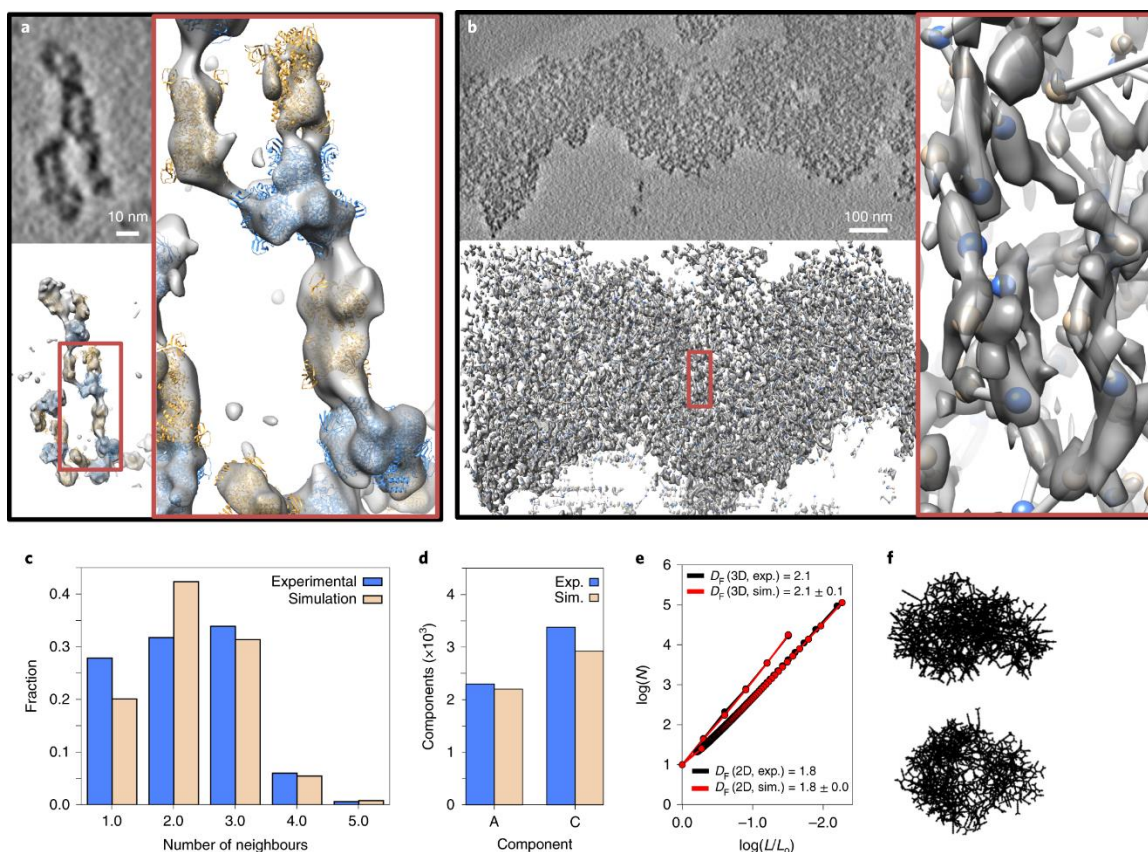
**a–d**, Longer fractal-like structures and branch-like and flower-like structures are seen in helium ion microscopy (HIM) (**a–c**) and AFM (**d**). **e**, Representative HIM images for assemblies obtained at different concentrations of pY-AtzM1 (250 nM–3 μM) while maintaining a fixed concentration of AtzCM1-SH2 (2 μM). Increasing concentrations of pY-AtzM1 result in larger assemblies with higher fractal dimensions. **f**,  $D_F$  and  $\lambda$  (the fractal dimension and lacunarity of the images) are similar for images obtained from different microscopy techniques. **g,h**, HIM images show fractal-like assembly formation with pY-AtzM1 and AtzCM1-SH2 (**g**), while the Gly-Ser-rich linker-containing variants form globular assemblies under these conditions (**h**).

Fractal patterns were not observed at any component stoichiometry without the addition of ATP and Src kinase, with unphosphorylated proteins or on drying the buffer (to preclude precipitation-induced assembly formation by the salt in the buffer), demonstrating that fractal structures are formed by designed components (Supplementary Fig. [23](#)). Similarly, fractal topologies were not detected when long ((GSS)<sub>10</sub>), conformationally flexible Gly-Ser-rich linkers were used to fuse the SH2 domain and pY tag to AtzC and AtzA, respectively. In mixtures of these proteins, a densely packed globular topology was detected with HIM, typical of amorphous precipitates (Fig. [3g,h](#) and Supplementary Fig. [24](#)). Thus, the surface-induced patterns observed with designed AtzC and AtzA are selectively formed following inter-component association in the designed geometries but not upon isotropic, random association, as expected for the highly flexible Gly-Ser-rich linker-containing variants.

#### 8.4.4 Structural characterization of assemblies using cryo-electron tomography

Transmission electron microscopy (TEM) of designed AtzA-AtzC proteins also revealed branching, dendritic networks reminiscent of fractal intermediates observed in biosilica formation<sup>14</sup> (Supplementary Fig. [25](#)). To further investigate the conformations of designed assemblies in solution and to obtain sufficiently high-resolution structures to test the validity of our design approach, we characterized the assemblies using cryo-electron tomography (cryo-ET; Fig. [4](#) and Supplementary Videos [2](#) and [3](#)). Assemblies generated by mixing 3  $\mu$ M pY-AtzA and 2  $\mu$ M AtzC-SH2 (or corresponding AtzA and AtzC fusions with Gly-Ser-rich linkers as controls) were blotted on a grid, frozen and visualized on a cryo-electron microscope. Due to the increased image contrast from Volt

phase plates in our microscope set-up, pY-AtzA and AtzC-SH2 complexes in assembly tomograms were easily identified as density clusters. In contrast, constructs with Gly-Ser-rich linkers connecting the pY and SH2 domain with AtzA and AtzC did not form porous clusters but instead (~90% of the sample) formed large, dense globular clumps (Supplementary Fig. 26) where individual components were not resolvable. These large topology changes on the micrometre scale (as observed by both cryo-ET and HIM) upon conformational flexibility changes at the nanometre scale further reinforce the importance of directional association in our modular fractal assembly design framework.



**Figure 94-8.** Fig. 4: Assembly characterization with cryo-ET.

**a,b**, Observed topologies in solution for a small (**a**) and large (**b**) assembly, in which the subtomograms were extracted and fit with Rosetta models. For the small assemblies, atomic-resolution models of pY-AtzAM1 (blue) and AtzCM1-

SH2 (tan) were fitted to reveal the intercomponent connections along assembly branches. For the large assembly, due to the lower resolvability in this region of the sample, only geometric centres of density were used to assign to individual components (blue and tan spheres; [Supplementary Section 3.6](#)). **c**, Spatially proximal neighbour distribution from the cryo-ET-derived images compared to simulated assemblies. **d**, Relative component distribution in the cryo-ET image and from simulations. **e,f**, Image analysis (2D), using a box counting method, of the cryo-ET tomography subtomograms converted into 2D projections shows a similar fractal dimension. The standard deviation in the simulation is calculated from 100 simulations composed of ~5,000 components. 3D box counting revealed a similar fractal dimension (slope) between the experimentally observed and simulated assemblies. Parameters for the simulation that match the experimental data are  $P_{\text{term}}=0.1$ ,  $C_{\text{frac}}=1.0$  and  $kT=9.0$ . Two representative 2D projections with the matching parameters are shown (**f**).

Computational annotation of the density clusters formed by designed components in cryo-ET-derived images was performed based on individual molecular envelopes of components derived from Rosetta models of pY-AtzA and AtzC-SH2, respectively, to identify inter-component connections along assembly branches (Fig. [4a](#)). The topology of the largest, nearly fully interconnected assembly based on electron density (Fig. [4b](#)), consisting of ~6,000 individual protein components, was further analysed and compared with an ensemble of simulated structures with approximately the same number of components. We compared the observed distributions of nearest-neighbour counts for AtzA-pY (Fig. [4c](#) and [Supplementary Fig. 27](#)), relative numbers of component types incorporated (Fig. [4d](#)), angular distribution (C-A-C connections, [Supplementary Fig. 28](#)) and the observed fractal dimension (Fig. [4e](#)) of the assemblies with ensembles of structures generated using computational modelling (Fig. [4f](#)) and found good agreement between the data and our simulations performed at specific parameter values ( $P_{\text{term}}=0.1$ ,  $C_{\text{frac}}=1.0$ ,  $kT=9.0$ ). The observed nearest-neighbour distribution for the AtzA-pY component shows that a large majority of these proteins are connected to one, two or

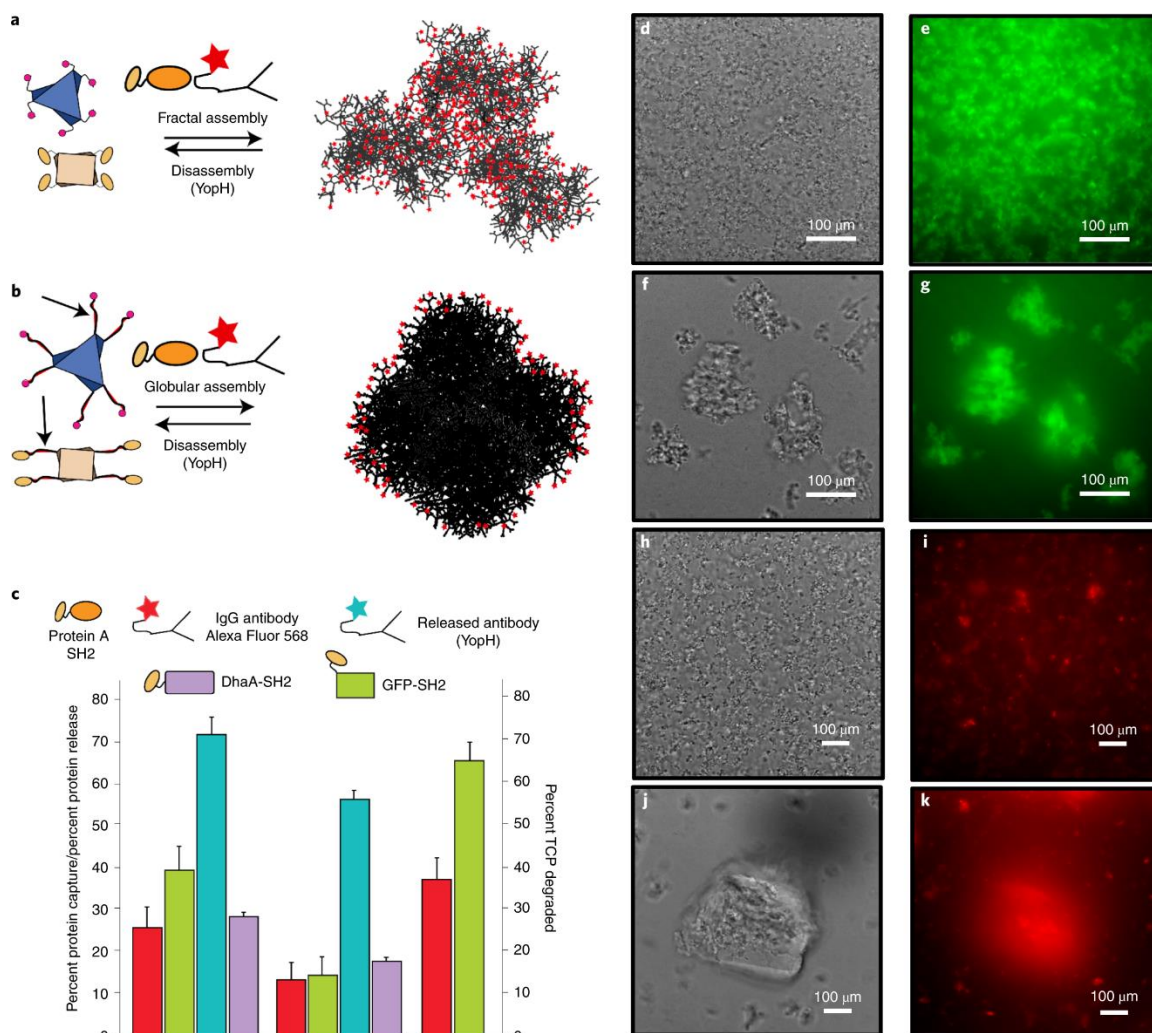
three neighbouring AtzC-SH2, in agreement with the divalent connections envisioned in the design model and implemented in the simulated assemblies (Fig. 1). Additionally, a small but significant number of AtzA-pY proteins have four AtzC neighbours in both the computational ensemble and the cryo-ET images, which indicates physically unconnected components being proximal to each other in space due to the packing in the assembly, although a small number of monovalent connections cannot be definitively ruled out in the cryo-ET images (Fig. 4c). We found that the fractal dimensions from the cryo-ET images and simulations (2.1) show good agreement (Fig. 4e). The expected fractal dimension for a DLA-like cluster, which results from isotropic interactions, is 2.3, and the observed decreased fractal dimension (2.1) indicates the anisotropic nature<sup>32,33,34</sup> of the underlying protein–protein interactions as encoded in the design approach (Fig. 1). Particle counting (and volume estimation) in a convex hull enclosing the largest assembly component yields an approximate local concentration of the proteins as ~600–700  $\mu\text{M}$ , an ~125-fold increase compared to their bulk concentration (3  $\mu\text{M}$  AtzA-pY and 2  $\mu\text{M}$  AtzC-SH2). The particle density in the fractal assembly is ~70,000 particles  $\mu\text{m}^{-3}$  whereas the calculated densities of 2D and 3D crystalline lattices of similar volumes are estimated to be ~4,000 and ~40,000 particles  $\mu\text{m}^{-3}$  (Supplementary Fig. 1). The high particle density in the fractal while maintaining porosity leads to a high effective surface area, a characteristic feature of macroscopic fractal objects such as trees and sponges. Although we could assign individual density clusters to individual components, the thickness of the ice in this region of the sample lowers resolvability and precludes direct measurement of orientation of the AtzA-pY and AtzC-SH2 components with respect to each other for comparison with Rosetta-calculated landscapes (Fig. 1). Although there is significant

heterogeneity in assembly sizes (~60% of the proteins adsorbed on the cryo-ET grid are parts of smaller assemblies) and topologies (Supplementary Fig. [29](#)), the observed increase in the effective concentrations concomitant with a high effective surface area with numerous solvent channels (Fig. [4a–f](#)) indicates that induced fractal-like structure formation is a viable strategy to engineer protein assemblies with favourable sponge-like properties.

#### 8.4.5 Functional characterization of designed assemblies

We next investigated if the observed textured, sponge-like topology, resulting in a high surface area:volume ratio in the fractal assembly, endows it with similar enhanced material capture ('soaking up') properties on the nanoscale as observed for macroscopic sponges. We reasoned that the lacunarity ('gappiness') of the fractal structure and use of an excess AtzA-pY component under fractal-forming stoichiometries would lead to several phosphopeptide sites on AtzA being open and accessible. The observed large pore sizes (Fig. [4b](#)) would enable access to these sites for molecular capture of nanometre-sized, macromolecular moieties bearing SH2 domains. In contrast, due to their dense, globular structure, amorphous assemblies generated with Gly-Ser-rich linker-containing components would have less available binding sites, resulting in a lower loading capacity (Supplementary Figs. [24](#) and [26](#)). To test the molecular capture properties of assemblies, we first used two fusion proteins in which macromolecular cargo proteins were fused to an SH2 domain: SH2-GFP, SH2-DhaA (an engineered DhaA enzyme for the degradation of the groundwater pollutant 1,2,3-trichloropropane, TCP), and measured the amount of cargo proteins captured by fractal and globular assemblies generated using identical

amounts of component proteins (Fig. [5](#)). Indeed, fractal assemblies captured greater amounts of cargo, as evidenced by fluorescence (green fluorescent protein, GFP) and enzymatic activity (DhaA) measurements, respectively (Fig. [5](#)). Fluorescence microscopy of SH2-GFP containing assemblies revealed that, as anticipated from cryo-ET studies, the immobilized cargo protein was distributed throughout the assembly and localized to the surface for fractal and globular assemblies, respectively (Fig. [5d–g](#)). To develop a more broadly applicable approach for exploiting the efficient molecular capture properties of fractal assemblies, we generated and utilized a SH2-protein A fusion protein to capture a fluorescent immunoglobulin-G (IgG) antibody. As observed for SH2-GFP and SH2-DhaA, fractal assemblies can efficiently capture this antibody (Fig. [5h–k](#) and Supplementary Fig. [30](#)). Furthermore, incubation of antibody-loaded assemblies with YopH phosphatase enzyme permits release of captured cargo antibodies (Fig. [5](#)). As all full-length IgG antibodies universally have the binding sites for protein A (their Fc-domains), antibody-loaded fractal assemblies should enable (1) efficient molecular capture of a variety of macromolecular and small-molecule antigens and (2) phosphorylation-dependent antibody purification<sup>[55,56,57](#)</sup>.



**Figure 95-8.** Fig. 5: Fractal assemblies capture and release greater amounts of cargo compared to globular assemblies.

**a,b**, Scheme for the envisioned reversible capture of cargo proteins for fractal (**a**) and globular (**b**) structures. Red stars denote an example cargo protein (antibody). **c**, Percent protein capture was measured for the 3:2 fractal (assemblies obtained with 3 AtzA-pY: 2 AtzC-SH2), 3:2 Gly-Ser (GS) linker (globular assemblies obtained at the same stoichiometry with fusion proteins containing long GS-rich linkers) and 3:1 fractal (assemblies obtained with 3 AtzA-pY: 1 AtzC-SH2). The 3:2 fractal captured more IgG antibody (red bars) and GFP-SH2 (green bars), and degraded more substrate TCP (purple bar; reflecting the higher capture efficiency of enzyme DhaA-SH2) than the 3:2 GS linker assemblies. In addition, the 3:2 fractal released more captured antibody compared to the 3:2 GS linker when incubated with YopH phosphatase (blue bars). The plotted error is the standard deviation of the percent protein capture/release/TCP degradation obtained from three independent measurements. The error used was a standard deviation of the percent capture/release/TCP degradation over the triplicates. **d-g**, Confocal fluorescence microscopy images of the three-component assembly with

GFP-SH2 showing the topology of incorporation of GFP-SH2 into the fractal assembly (**d,e**) and the incorporation of GFP-SH2 into the globular assembly (**f,g**). **h–k**, IgG antibody Alexa Fluor 568 incorporation into the fractal assembly (**h,i**) and incorporation into the globular assembly (**j,k**).

In our design framework, fractal loading capacity is determined by the number and accessibility of open phosphopeptide binding sites in the assembly. Thus, assemblies formed by 3 (AtzA-pY):1 (SH2-AtzC) are expected to have a greater loading capacity than those formed by 3 (AtzA-pY):2 (SH2-AtzC). Indeed, as anticipated, more antibody was captured and released by the former than the latter (Fig. [5c](#)), demonstrating that customized optimization of molecular capture-and-release of specific nanoscale objects should be possible by varying the component stoichiometry to obtain the fractal properties on nanometre to micrometre scales. Finally, we asked if the observed functional advantages of a fractal topology over a globular one would extend to the capture and transport of small molecules within the assembly by measuring the efficacy of atrazine degradation. As cargo we incorporated AtzB—the third pathway enzyme (apart from AtzA and AtzC) required to convert atrazine to the relatively benign metabolite cyanuric acid (Supplementary Figs. [31](#) to [34](#)). Although both the fractal and globular assemblies appear to be more robustly active under harsh reactions compared to unassembled enzymes (Supplementary Fig. [35](#)), and when immobilized on a Basotect polymer foam (Supplementary Fig. [36](#)), both globular and fractal assemblies are equally active (Supplementary Fig. [37](#)). The significantly small size of atrazine (radius of gyration,  $R_g < 1$  nm) and other metabolic pathway intermediates probably allows them to diffuse equally efficiently in either assembly as the smaller solvent channel size in the globular assembly may not be an impediment for a small molecule guest as opposed to

macromolecular guest molecules. Constructing fractal-like shapes with smaller sized proteins may allow access to smaller solvent channels.

## 8.5 Discussion

Our results demonstrate an approach by which fusion proteins may be designed to self-assemble into fractal-like morphologies on the 10 nm–10  $\mu$ m length scale. The design strategy is conceptually simple, modular and should be applicable to any set of oligomeric proteins featuring cyclic, dihedral and other symmetries, such that multivalent connections, anisotropy and geometric degeneracy of binding can be used to controllably generate a broad range of sizes and morphologies of fractal shapes with proteins. In contrast with computational design of integer-dimensional protein assemblies where considerable remodelling of protein surfaces is necessary to meet the exacting geometric requirements for inter-component binding<sup>24,25</sup>, our design approach to obtain fractal-like morphologies involves few substitutions on protein surfaces. Instead, design goals are encoding high affinity via fusion of binding domains, and binding anisotropy and geometric degeneracy via short, flexible loops (see [Supplementary Discussion](#)). Although we used SH2 domain-pY peptide fusions as the high-affinity modular connecting elements to endow phosphorylation responsiveness, the same design strategy should be applicable for the incorporation of other peptide recognition domains, responsive to other chemical or physical stimuli. The combination of multivalency and chain flexibility is a key determinant of other recently discovered phases formed by proteins, including droplets formed by liquid–liquid phase separation<sup>58</sup>. Our results show that this rich phase behaviour of proteins<sup>44</sup> also includes fractal-like morphologies that form colloidal particles with constituent microscopic molecular networks which may be visualized at

high resolution using cryo-ET. Given the wide-ranging applications of fractal-like nanomaterials for molecular capture, further development in the design of protein-based fractals described here is expected to enable the production of novel classes of bionanomaterials and devices.

## 8.6 References

1. Mandelbrot BB: *The Fractal Geometry of Nature*. (W. H. Freeman & Company, 1982).
2. Stanley HE, Meakin P: Multifractal **Phenomena in Physics and Chemistry**. *Nature* 1998, **335**: 405-409.
3. Losa, GA: *Fractals in biology and medicine. Volume IV*. (Birkhäuser, 2005).
4. Fairbanks MS et al.: **Fractal electronic devices: simulation and implementation**. *Nanotechnology* 2011, **22**.
5. Soleymani L, Fang ZC, Sargent EH & Kelley SO: **Programming the detection limits of biosensors through controlled nanostructuring**. *Nat Nanotechnol* 2009, **4**: 844-848.
6. Ge J, Lei JD, & Zare RN: **Protein-inorganic hybrid nanoflowers**. *Nat Nanotechnol* 2012, **7**: 428-432.
7. Zhang PC & WangST: **Designing Fractal Nanostructured Biointerfaces for Biomedical Applications**. *Chemphyschem* 2014, **15**:1550-1561
8. Lim B et al.: **Pd-Pt Bimetallic Nanodendrites with High Activity for Oxygen Reduction**. *Science* 2009 **324**:1302-1305.
9. Cerofolini GF, Narducci D, Amato P & Romano E: **Fractal nanotechnology**. *Nanoscale Res Lett* 2008, **3**: 381-385.

10. Newkome GR et al.: **Nanoassembly of a fractal polymer: A molecular "Sierpinski hexagonal gasket"**. *Science* 2006, **312**:1782-1785.
11. Shang J, et al.: **Assembling molecular Sierpinski triangle fractals**. *Nat Chem* 2015, **7**:389-393.
12. Newkome GR, Moorefield CN: **From 1 -> 3 dendritic designs to fractal supramacromolecular constructs: understanding the pathway to the Sierpinski gasket**. *Chem Soc Rev* 2015, **44**; 3954-3967
13. Shin S et al.: **Polymer Self-Assembly into Unique Fractal Nanostructures in Solution by a One-Shot Synthetic Procedure**. *J Am Chem Soc* 2018, **140**.
14. Tikhomirov G, Petersen P, & Qian L: **Fractal assembly of micrometre-scale DNA origami arrays with arbitrary patterns**. *Nature* 2017, **552**: 67-71.
15. Zhang F, Nangreave J, Liu Y & Yan H: **Reconfigurable DNA origami to generate quasifractal patterns**. *Nano Lett* 2012, **12**: 3290-3295.
16. Rothmund PW, Papadakis N & Winfree E: **Algorithmic self-assembly of DNA Sierpinski triangles**. *PLoS Biol* 2004, **2**.
17. Astier Y, Bayley H & Howorka S: **Protein components for nanodevices**. *Curr Opin Chem Biol* 2005, **9**:576-584.
18. Murr MM & Morse DE: **Fractal intermediates in the self-assembly of silicatein filaments**. *P Natl Acad Sci USA* 2005, **102**:11657-11662.
19. Khire TS, Kundu J, Kundu SC & Yadavalli VK: **The fractal self-assembly of the silk protein sericin**. *Soft Matter* 2010, **6**:2066-2071.
20. Lomander A, Hwang WM, & Zhang SG: **Hierarchical self-assembly of a coiled-coil peptide into fractal structure**. *Nano Lett* 2005, **5**:1255-1260.

21. Shen W, Lammertink RG, Sakata JK, Kornfield JA & Tirrell DA: **Assembly of an artificial protein hydrogel through leucine zipper aggregation and disulfide bond formation.** *Macromolecules* 2005, **38**: 3909-3916.
22. Li B, et al.: **Nonequilibrium Self-Assembly of pi-Conjugated Oligopeptides in Solution.** *ACS Appl Mater Interfaces* 2017 **9**: 3977-3984.
23. McManus JJ, Charbonneau P, Zaccarelli E, & Asherie N: **The physics of protein self-assembly.** *Curr Opin Colloid In* 2016, **22**: 73-79.
24. King NP et al.: **Computational design of self-assembling protein nanomaterials with atomic level accuracy.** *Science* 2012, **336**: 1171-1174.
25. Hsia Y et al.: **Design of a hyperstable 60-subunit protein dodecahedron.** *Nature* 2016, **535**: 136-139.
26. Suzuki Y et al.: **Self-assembly of coherently dynamic, auxetic, two-dimensional protein crystals.** *Nature* 2016, **533**: 369-373.
27. Sinclair JC, Davies KM, Venien-Bryan C & Noble ME: **Generation of protein lattices by fusing proteins with matching rotational symmetry.** *Nat Nanotechnol* 2011, **6**:558-562.
28. Padilla JE, Colovos C & Yeates TO: **Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments.** *Proc Natl Acad Sci USA* 2001, **98**: 2217-2221.
29. Zhang J, Zheng F & Grigoryan G: **Design and designability of protein-based assemblies.** *Curr Opin Struc Biol* 2014, **27**: 79-86.
30. Subramanian RH et al.: **Self-Assembly of a Designed Nucleoprotein Architecture through Multimodal Interactions.** *ACS Cent Sci* 2018, **4**: 1578-1586.

31. Churchfield LA & Tezcan FA: **Design and Construction of Functional Supramolecular Metalloprotein Assemblies.** *Acc Chem Res* 2019.
32. Sontz PA, Song WJ & Tezcan FA: **Interfacial metal coordination in engineered protein and peptide assemblies.** *Curr Opin Chem Biol* 2014, **19**:42-49.
33. Brodin JD et al.: **Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays.** *Nat Chem* 2012, **4**:375-382.
34. Ringler P & Schulz GE: **Self-assembly of proteins into designed networks.** *Science* 2003, **302**: 106-109.
35. Lindenmayer A: **Mathematical Models for Cellular Interactions in Development .2. Simple and Branching Filaments with 2-Sided Inputs.** *J Theor Biol* 1968, **18**.
36. Lindenmayer A: **Mathematical Models for Cellular Interactions in Development .I. Filaments with 1-Sided Inputs.** *J Theor Biol* 1980, **18**, 280.
37. Glotzer SC & Solomon MJ: **Anisotropy of building blocks and their assembly into complex structures.** *Nat Mater* 2007, **6**: 557-562.
38. Zhang Z & Glotzer SC: **Self-Assembly of Patchy Particles.** *Nano Lett* 2004, **4**: 1407-1413.
39. Kartha MJ & Sayeed A: Phase transition in diffusion limited aggregation with patchy particles in two dimensions. *Phys Lett A* 2016, **380**: 2791-2795.
40. Nicolas-Carlock JR, Carrillo-Estrada JL & Dossetti V: **Fractality a la carte: a general particle aggregation model.** *Sci Rep-Uk* 2016, **6**.

41. Guesnet E, Dendievel R, Jauffres D, Martin CL & Yrieix B: **A growth model for the generation of particle aggregates with tunable fractal dimension.** *Physica A* 2019, **513**: 63-73.
42. Mansbach RA & Ferguson AL: **Patchy Particle Model of the Hierarchical Self-Assembly of pi-Conjugated Optoelectronic Peptides.** *J Phys Chem B* 2018, **122**: 10219-10236.
43. Bianchi E, Tartaglia P, Zaccarelli E & Sciortino F: **Theoretical and numerical study of the phase diagram of patchy colloids: ordered and disordered patch arrangements.** *J Chem Phys* 2008, **128**.
44. Lomakin A, Asherie N & Benedek GB: **Aeolotopic interactions of globular proteins.** *Proc Natl Acad Sci USA* 1999, **96**: 9465-9468.
45. Vacha R & Frenkel D: **Relation between molecular shape and the morphology of self-assembling aggregates: a simulation study.** *Biophys J* 2011, **101**: 1432-1439.
46. Bianchi E, Tartaglia P, La Nave E & Sciortino F: **Fully solvable equilibrium self-assembly process: fine-tuning the clusters size and the connectivity in patchy particle systems.** *J Phys Chem B* 2007, **111**: 11765-11769.
47. Yan Y, Huang J & Tang BZ: **Kinetic trapping - a strategy for directing the self-assembly of unique functional nanostructures.** *Chem Commun (Camb)* 2016, **52**: 11870-11884.
48. Wackett LP, Sadowsky MJ, Martinez B & Shapir N: **Biodegradation of atrazine and related s-triazine compounds: from enzymes to field studies.** *Appl Microbiol Biotechnol* 2002, **58**: 39-45.

49. Kaneko T et al: **Superbinder SH2 domains act as antagonists of cell signaling.** *Sci Signal* 2012, **5**.
50. Yang L et al.: **Computation-Guided Design of a Stimulus-Responsive Multienzyme Supramolecular Assembly.** *Chembiochem* 2017, **18**: 2000-2006.
51. Das R & Baker D: **Macromolecular modeling with rosetta.** *Annu Rev Biochem* 2008, **77**; 363-382.
52. Pellegrini M, Wukovitz SW & Yeates TO: **Simulation of protein crystal nucleation.** *Proteins* 1997, **28**: 515-521.
53. Masters BR: **Fractal analysis of the vascular tree in the human retina.** *Annu Rev Biomed Eng* 2004, **6**:427-452.
54. Witten TA. & Sander LM: **Diffusion-Limited Aggregation, a Kinetic Critical Phenomenon.** *Phys Rev Lett* 1981, **47**: 1400-1403/
55. Swartz AR & Chen W: **SpyTag/SpyCatcher Functionalization of E2 Nanocages with Stimuli-Responsive Z-ELP Affinity Domains for Tunable Monoclonal Antibody Binding and Precipitation Properties.** *Bioconjug Chem* 2018, **29**: 3113-3120.
56. Bilgicer B et al.: **A non-chromatographic method for the purification of a bivalently active monoclonal IgG antibody from biological fluids.** *J Am Chem Soc* 2009, **131**: 9361-9367.
57. Handlogten MW, Stefanick JF, Deak PE, & Bilgicer B: **Affinity-based precipitation via a bivalent peptidic hapten for the purification of monoclonal antibodies.** *Analyst* 2014, **139**: 4247-4255.

58. Brangwynne CP, Tompa P & Pappu RV: **Polymer physics of intracellular phase transitions**. *Nat Phys* 2014, **11**: 899-904.

## 8.7 Supplemental Information

### 8.7.1 SI 1. Computational methods

**(SI 1.1) Preparation of scaffolds** – Crystal structure files for AtzA (PDB:4V1X) and AtzC (PDB:2QT3) were subject to several preparatory scripts to clean, symmetrize, and process the files for Rosetta Design<sup>1-3</sup>. The processed crystal structure files were then subject to a Rosetta Fast Relax<sup>4</sup> protocol to obtain starting structures of sufficiently low Rosetta Energy to serve as starting structures. To prepare a library of conformations predicted to propagate into a fractal structure, we first aligned the proteins along paired C<sub>2</sub> symmetry axes (A+B chains for both AtzA and AtzC). We then translated AtzC along the aligned C<sub>2</sub> symmetry axis until the backbone (N, C, Ca, O) atoms of each structure were at least 3Å apart to find the minimum starting distance between the centers of mass (125Å). From the minimum starting distance we translated AtzC(monomer) in intervals of 1Å to a maximum distance of 145Å. For each translated AtzC(monomer) position we rotated the AtzC(monomer) about the C<sub>2</sub> symmetry axis at a core set of angles i.e.,  $180 \pm (0, 35.25, 54.75, 90, 125.25, 144.75, \text{ and } 180)$ . As detailed below, we used this scaffold library to stabilize simultaneously a subset of rotations about the paired C-symmetric axes known to favor 2D or 3D crystal geometries (Fig. S1). These values were chosen as they were expected to allow propagation of the assembly in a stochastic manner – we hypothesized that a mixture of connections at a set of angles should result in a fractal assembly (Fig. S1).

### (SI 1.2) RosettaMatch: simultaneous fusion domain and peptide pair stitching –

After visual inspection of the two-component scaffold placements, we noted the accessibility of the AtzA N-terminus and the AtzC C-terminus along the  $C_2$  symmetry axis (chains A+B). Therefore, we decided to fuse the N-terminus of an fyn-SH2 super-binder (PDB:1A0T) to the C-terminus of AtzC and the C-terminus of the fyn-SH2 peptide binding partner to the N-terminus of AtzA. To achieve simultaneous fusion, we converted the SH2-peptide crystal structure into an all- $C\alpha$  'ligand' file and used RosettaMatch<sup>5</sup> (Fig. S2) with geometric constraints to sample all sterically feasible rigid body placements of the SH2-peptide between each AtzA-AtzC pair in the two-component scaffold library. RosettaMatch requires a set of 6 geometric degrees of freedom, imposed as constraints in order to orient a given "ligand" (SH2-peptide  $Ca$  trace) model with respect to each desired amino acid contact. Geometric constraints used to coordinate the SH2 domain-peptide complex for simultaneous fusion were derived from backbone atom positions and orientations using a non-redundant protein library generated by the RCSB-PDB<sup>6</sup>. From N to C terminus, regardless of secondary structure, we collected distances and angles between backbone atoms ( $C\alpha$ , nitrogen, and carboxyl carbon) up to and including 7 residues downstream (sequence-space) of each residue along the primary structure. The averages and standard deviations of these distributions were used to place geometric constraints between residues of the AtzA-AtzC termini and the all- $C\alpha$  SH2-peptide ligand to force only geometrically allowable backbone fusions. The full-atom SH2-peptide crystal structure was re-threaded back onto each of the matched SH2-peptide ligands creating 7,005 models with paired termini in proximally close and geometrically favorable positions. Rosetta GeneralizedKIC (kinematic loop closure)<sup>7</sup> was

used to covalently link the paired termini and generate 3 potential linker-models for each matched SH2-peptide model, creating a library of 21,015 fused and bound AtzA-AtzC pairs. Geometric constraint files and GenKIC XML protocol files are provided in a supplementary data zipped file.

**(SI 1.3) Rosetta Design: interface design** –A Rosetta-based design protocol was used to stabilize the novel interfaces formed from the placements of SH2-peptide and protein monomers obtained using RosettaMatch (followed by loop closure) as described in SI 1.2. In the design protocol we allowed linker residues ( $\pm 4$  residues around the fusion site) and AtzC-SH2 domain interface residues ( $C\alpha$ - $C\alpha$  distance  $< 6\text{\AA}$ ) to change residue identity. All backbone atoms with the exception of the linker residues were constrained with atom-coordinate constraints to favor the SH2-peptide placements determined in the RosettaMatch step. Mutations at the linker region,  $\pm 4$  residues around the fusion site, which alleviated steric clashes with the backbone or sidechains of AtzC/SH2 domain were accepted such that they stabilized multiple intercomponent angles of attachment (Table S1). All other mutations were reverted to native residue identities before a subsequent round of repack and energy minimization<sup>8</sup>. Based on our calculations, we selected 5 variants of AtzA fused to the SH2-peptide recognition sequence and 5 variants of the AtzC fused to the SH2 domain for experimental characterization. Table S1 shows a list of mutations and the value of rotation stabilized for each design. Rosetta command lines and XML files used for design are provided in supplementary zipped file.

**(SI 1.4) Generation of energy landscape** – To evaluate the energy landscape of the designed component pair (pY-AtzAM1 and AtzC-SH2M1) along the symmetrically aligned rotation-translation degrees of freedom we performed a Rosetta Symmetric

FastRelax protocol on conformations of AtzA-AtzC pairs. Each conformation was represented by three parameters, translation ( $d$ ), rotation ( $\theta$ ), and axis-binding preference (vertex or edge centered). Parameter values  $d$  in range 120Å to 145Å in steps of 1Å;  $\theta$  in range 0° to 360° in steps of 5° were used to identify their predicted preferred binding modes. We generated this energy profile and plot in (Figure 1E-F) conformations whose evaluated binding energy scored better than the wild-type components (504 models). This binding energy was calculated by separating the binding components to a distance of >500 Å and repacking the components. Binding energy values were used to compute a Boltzmann-weighted probability distribution used in coarse grained simulations described below. The Rosetta command lines and XML files for energy landscape generation are provided in supplementary zipped file.

#### **(SI 1.5) Coarse-graining AtzA-C oligomers for stochastic fractal growth simulations**

—we used a coarse-grained representation of our symmetric oligomers by reducing each chain to 10 representative points in space (60 and 40 for whole hexamer and tetramer respectively). To coarse-grain we used a K-means-like clustering algorithm to place the 10 points at locations with the highest concentration of C $\alpha$  atoms in each monomer (chain A). We then calculated and applied the symmetric transform to the 10 representative points to obtain a coarse-grained representation of each oligomer (hexamer and tetramer). When each representative point is converted to a sphere with a 12Å radius, the coarse-grained model effectively mimics the overall shape and size of the full-atom model.

**(SI 1.6) Stochastic fractal assembly simulation**— In order to predict the supramolecular structure and topology we developed a stochastic fractal assembly simulation protocol

that utilizes Boltzmann weighted probability distributions for an ensemble of predicted low-energy binding modes along the  $C_2$ -symmetry axes of the AtzA-AtzC pairs as described in SI 1.5. The algorithm operates by starting with one oligomer (AtzA for this study) and attaches each complementary oligomer layer-by-layer. The Boltzmann probability distribution was used to decide how the oligomers in each layer were placed. A few key assumptions were made during the simulations. We assumed: 1) The symmetric divalent connection along a  $C_2$ -symmetry axis (two chains of pY-AtzA bound two chains of AtzC-SH2) would be energetically more likely than the monovalent connection formed between just one chain from each oligomer—reducing the probability of monovalent connection to an insignificant value. 2) Flexibility in the linker region would only lead to variations along the  $C_2$ -symmetry axis via the translation and rotation parameters used in design—maintaining the inherent symmetry found in either oligomer. 3) Mixed vertex-centered and edge-centered species could occur around a single AtzA. This would lead to a substructure where two AtzC oligomers have a  $180^\circ$ -angle about an AtzA component, different from the  $120^\circ$ -angles when pure edge-centred and vertex-centered binding geometries are considered. 4) Changes in size and topology would arise from concentration changes of the enzymes and would need to be represented in the algorithm. 5) During fractal growth it is possible (and likely) that oligomers in one layer could come within  $125\text{\AA}$  (minimum connected distance) of other oligomers within another layer even if they are not directly connected. The details of this algorithm are described below.

Energy landscapes calculated in SI 1.4 were used to stochastically propagate the coarse grained A-C components during simulation. We varied the  $kT$  term to obtain a total of 5

different Boltzmann weighted probability distributions ( $kT = 1, 3, 5, 7$ , and  $9$ ).

Propagation was achieved by alternating layers of AtzA and AtzC components starting from an initial seed component (pY-AtzA in this study) which would continue until either placement of new components was determined either impossible or improbable or an external criterion was met (number of layers, size of particle, etc.). The propagation algorithm involves 6 steps at any given layer:

1) Using  $C_{\text{frac}}$ , randomly select a fraction of the components from the previous layer (or seed if 1st iteration) to continue propagation. Components not chosen will cease to propagate for the remainder of the simulation. For example, if  $C_{\text{frac}} = 0.33$ , a third of the available connection points are considered propagatable.  $C_{\text{frac}}$  models the relative stoichiometry of the two components.

2) Iterate over the selected components determined in step 1 and:

2a) Randomly select an available  $C_2$ -symmetry axis of the individual selected in (2).

2b) Using a Boltzmann weighted probability distribution of possible  $d$ - $\theta$ -axis conformations, attempt to apply a component with an orientation at random along the  $C_2$  symmetric axis (2a).

2c) Choose whether or not to keep the selected  $C_2$ -symmetry axis (2b) based on a termination probability ( $P_{\text{term}}$ ). The termination probability models the binding affinity. For example, a weaker binding SH2 domain variant would have a higher  $P_{\text{term}}$  compared to a tight-binding variant.

2d.1) If (2c) passes the term, apply the rigid body transformation ( $d$  and  $\theta$ ) and append as a member of the next layer.

2d.2) If (2c) fails the term, mark the  $C_2$ -symmetry axis (2a) of the individual selected in (2) as unviable and continue.

3) Repeat 2a-d.1 until all  $C_2$ -symmetry axes of individual (2) are exhausted.

4) Perform a coarse grid-based clash check to ensure new layer members are sterically non-clashing with any components of the assembly.

5) Repeat Steps 2-4 until all of the components chosen in (1) are exhausted.

6) Move to the next layer.

**(SI 1.7) Temperature, fraction, and term parameter sweep** – Varying the fraction  $C_{\text{frac}}$

(1) and termination probability  $P_{\text{term}}$  (5-6b) parameters gave rise to changes in topology and structure. We created 100 fractal models for each combination of  $C_{\text{frac}}$  (range: 0.1-1.0, interval: 0.1) and  $P_{\text{term}}$  (range: 0.0-0.9, interval: 0.1) using the 5 different Boltzmann weighted probability distributions (with varying temperature)—creating 50,000 total fractal assemblies. These parameter sweep simulations were performed for 15 layers for each parameter combination. We analyzed each particle's individual size, number of layers, AtzA branch ratio (number of AtzC units bound to a unit of AtzA, lacunarity ( $\lambda$ ), and dimensionality ( $D_f$ ) from a 2D image. For every combination of temperature, fraction, and term we averaged the data across the 100 fractal assemblies. The results can be found in Figure S5 and S6. In the range  $C_{\text{frac}}$  values 0.5-1 and  $P_{\text{term}}$  values 0.0-0.4, particle diameter, branch ratio, fractal dimension, and lacunarity increase with increasing

values of  $C_{\text{frac}}$  and decreasing values of  $P_{\text{term}}$ . Particle diameter and branch ratio decrease with increasing values of  $kT$  while lacunarity increases. Fractal dimension first increases until  $kT = 5$  and then decreases. In this range the number of layers are shown to remain the same.

**(SI 1.8) Preparing fractal models for image analysis** – Each coarse-grained assembly model obtained above was analyzed using a PyMOL script that would color the assembly components black, convert the background to white, show as spheres of scale  $12\text{\AA}$ , orient the image such that the longest diameters are in the X-Y plane, remove the glossy lighting and shine from the sphere models, and finally ray-trace render the image. This PyMOL script is provided in supplementary zipped file.

**(SI 1.9) Preparing helium ion microscopy (HIM) images for image analysis** – HIM images were loaded into ImageJ<sup>9</sup>. The initial image contrast was enhanced with 5-20% saturated pixels setting; this can be achieved with Process -> Enhance Contrast. We then create a new blank (black) image with the same pixel dimensions as the HIM image. Gaussian noise is added to the blank image with a standard deviation 5-10 (Process -> Noise -> Add Specified Noise). Background noise is subtracted from the HIM image using the noisy blank image (Process -> Image Calculator -> set Image1 to HIM image and image2 to noisy blank -> set operation to subtract). Finally, we create a binary image from the processed HIM image with subtracted background. The resulting image contains white protein islands on a black background. Individual fractal islands are then copy/pasted into a new blank (black) image using the polygon selection tool and are ready for fractal analysis.

**(SI 1.10) Determining fractal lacunarity and 2-D fractal dimension with ImageJ -**

The FracLac package<sup>10</sup> designed for ImageJ was used to determine both the 2D lacunarity and fractal dimension ( $D_f$ ). With FracLac mode on, outside of the standard parameters, we checked the 'alternate random generator' box and allowed the minimum pixel size to be 1, and the color code was turned off. We then ran in batch-mode to process all of the fractal images. ImageJ outputs four files: summary, box count per grid, scan types, and batch data. Lacunarity and dimension were taken from the summary file for the parameter sweep while the 2D log vs log plot values were taken from the box counting grid file ( $\epsilon$  and  $F$ ).

**(SI 1.11) Comparison of simulated and experimentally observed assemblies from Cryo-EM** – Fitting of the experimentally computed protein density (from Cryo-electron tomography) resulted in Cartesian coordinates representing the center of mass of the oligomeric components. To compare the experimental results to simulation we ran the simulation until at least a total of 5000 components were present in the model and calculated the geometric centers for all oligomeric components in the coarse-grained assembly to create new center-of-mass models. Using the experimentally derived Cartesian coordinates and the center-of-mass models we performed a computational analysis (SI 3.6) to evaluate the fractal size, nearest component neighbor distances, and relative AtzA-AtzC ratio (Fig. 3H,I). We analyzed the 3D fractal dimension (Fig. 3J) with a 3D box counting program that counts the number of geometric centers within a scaling (doubling) box size. The 2D fractal dimension (Fig. 3J) was calculated in the same way as previously mentioned (Fig. S4). We found highest agreement of simulations with  $kT = 9$ ,  $P_{\text{term}} = 0.1$ , and  $C_{\text{frac}} = 1.0$ . An array of fractal images that represent the average fractal for each value of  $P_{\text{term}}$  and  $C_{\text{frac}}$  at  $kT = 9$  can be found in Figure S6.

## SI 2. Experimental characterization methods

**(SI 2.1) Creation of the designed AtzA, AtzB, and AtzC fusion constructs** – The DNA sequence of the full-length *atzA* was amplified from the *pMD4::atzA*; *atzB* amplified from *pAAJLS3::atzB*; and *atzC* was amplified from *pKK223-3::atzC*.<sup>9–12</sup> The Src kinase activator phosphopeptide sequence, EPQYEEIPIYL, was created by ordering two complementary primers that formed a linear fragment encoding the peptide sequence, used with the amplified *atzA* gene and inserted into the linearized *pET15b+* vector through Gibson Assembly.<sup>13</sup> The Fyn SH2 superbinder gene was ordered as a gBlock fragment<sup>13,14</sup> and inserted into *pET29b+* (linearized with *NdeI* and *XhoI*) using Gibson Assembly. The Fyn SH2 amplified gene was designed to be placed on the C-terminal side of the *pET15b+::atzB* and *pET29b+::atzC* with a flexible GSS linker between the proteins. The Fyn SH2 superbinder amplified gene *SH2* and the *atzC* amplified gene were both inserted into the *pET29b+* linear vector using Gibson Assembly. The *atzBSH2* fusion gene was ordered as a Gibson fragment<sup>13</sup> and inserted into the *pET15b+* linear vector using Gibson Assembly. Point mutations were introduced using the QuickChange Site-Directed Directed Mutagenesis Kit (Agilent Technologies) to create the final designs for AtzA and AtzC models. DNA sequencing was used to confirm proper insertion and mutations (Genscript).

**(SI 2.2) AtzA and AtzC expression and purification** – The *pET15b+::atzApep* and *pET29b+::atzCSH2* plasmids were co-transformed into *Escherichia coli* BL21 (DE3) with *pAG* plasmid containing genes for the chaperone proteins, *groEL* and *groES*.<sup>15</sup> For

expression of the AtzA models a 10 mL LB culture with 30  $\mu\text{g/mL}$  of chloramphenicol and 100  $\mu\text{g/mL}$  of ampicillin was inoculated with a single colony and incubated overnight at 37°C and 250 rpm. For the expression of the AtzC models a 10 mL LB culture with 30  $\mu\text{g/mL}$  of chloramphenicol and 50  $\mu\text{g/mL}$  of kanamycin was inoculated. After growing overnight, the 10 mL cultures of the AtzA and AtzC models were used to inoculate 500 mL of LB media, which was grown at 37°C to an  $\text{OD}_{600}$  of 0.5-0.6, at which point the expression of chaperones was induced with the addition of 1% (wt/vol) L-arabinose and grown for an additional 1-2 hours at 16°C. Expression of the AtzA and AtzC models was then induced with 0.1mM IPTG (isopropyl- $\beta$ -D-thiogalactopyranoside) and grown overnight at 16°C. All subsequent steps were performed at 4°C. Cells were centrifuged at 6,000 x g for 30 min. Cell pellets were re-suspended in 30 mL of 25 mM HEPES, 200 mM NaCl, 5% glycerol, 40 mM imidazole, pH 7.5, and lysed by sonication. Cell extracts were obtained by centrifugation at 50,000 x g for 30 min at 4°C. Protein purification was performed using 5 mL Ni-NTA agarose resin (Qiagen) equilibrated with 10 mL of 25 mM HEPES, 200 mM NaCl, 5% glycerol, 40 mM imidazole, pH 7.5. The lysate was applied to the resin, the resin was washed with 45 mL of the same buffer, and the protein eluted with 20 mL of 25 mM HEPES, 200 mM NaCl, 5% glycerol, 400 mM imidazole, pH 7.5,. The purified protein was buffer exchanged (PD10-desalting column, GE Healthcare #17085101) into 50 mM HEPES, 100 mM NaCl, 5% glycerol, pH 7.4 (HNG). AtzA was expressed in high yields and precipitated if imidazole was not removed immediately after elution from the Ni-column. No precipitation was observed upon exchange of AtzA into HNG buffer and AtzC variants

did not precipitate in either buffer (for hours-days) at 4°C. Proteins were frozen using liquid nitrogen and stored at -80°C.

**(SI 2.3) AtzB expression and purification** – The *pET15b+::atzBSH2* plasmid was transformed into *E.coli* BL21 (DE3) cells. For expression of AtzB, a 10 mL LB culture with 100 µg/mL of ampicillin was inoculated overnight at 37°C and 250 rpm. The 10 mL overnight culture was used to inoculate 500 mL of LB media which was grown to an OD<sub>600</sub> of 0.5-0.7 and induced with 1 mM IPTG and grown overnight at 16°C. The same purification protocol for the AtzA and AtzC models was used for AtzB. AtzBSH2 did not express if grown with zinc sulfate, as had been done customarily in previous literature.<sup>11</sup>

**(SI 2.4) Src human kinase, super binder SH2 domain, SH2-DhaA expression and purification** – The expression plasmid for Src human kinase<sup>16</sup> (gift from John Chodera, Nicholas Levinson, and Markus Seeliger. Addgene plasmid # 79700) was co-transformed with the expression plasmid for *Yersinia* YopH protein tyrosine phosphatase (PTPase)<sup>16</sup> (gift from John Chodera, Nicholas Levinson, and Markus Seeliger, Addgene plasmid # 79749) into *E. coli* Rosetta2 (DE3) (Novagen). For Src kinase expression a 10 mL LB culture with 50 µg/mL spectinomycin and 100 µg/mL of ampicillin was inoculated with a single colony and incubated overnight at 37°C, 250 rpm. The overnight culture was used to inoculate 500 mL of LB media which was grown to an OD<sub>600</sub> of 0.5-0.7 and induced with 1mM IPTG and grown overnight at 18°C. The super binder SH2 domain and SH2-DhaA were transformed into *E. coli* BL21 (DE3) and expressed in the same way as the Src kinase above. Purification for the Src kinase was performed similarly and with the same buffers as AtzAM1, AtzBSH2, and AtzCM1. While, the super binder SH2 domain

and SH2-DhaA were purified with the same purification protocol but with the following buffers: a wash buffer containing 137 mM NaCl, 2.7 mM KCl, 10 mM Na<sub>2</sub>HPO<sub>4</sub>, 2 mM KH<sub>2</sub>PO<sub>4</sub>, pH 7.4, 20 mM imidazole and an elution buffer containing 137 mM NaCl, 2.7 mM KCl, 10 mM Na<sub>2</sub>HPO<sub>4</sub>, 2 mM KH<sub>2</sub>PO<sub>4</sub>, pH 7.4, 200 mM imidazole. All proteins were buffer exchanged into HNG, frozen in liquid nitrogen and stored at -80°C.

**(SI 2.5) YopH phosphatase construct, expression, and purification** – The linear catalytic domain *YopH* gene (residues 164-468) was amplified from *pET13S-A::YopH*<sup>16</sup> and inserted with Gibson Assembly into a linearized pET15b+ vector. A 10 mL LB culture with 100 µg/mL of ampicillin was inoculated with a single colony and incubated overnight at 37°C. The expression and purification protocol is the same as the protocol used for the Src kinase.

**(SI 2.6) Biuret hydrolase and cyanuric acid hydrolase expression and purification** – Biuret hydrolase (BH)<sup>17</sup> expression strain (*E. coli* DH5α) and the *Moorella* Cyanuric acid hydrolase (CAH)<sup>18</sup> strain (*E. coli* BL21 (DE3)) were provided by Dr. Larry Wackett. A 10 mL culture with 50 µg/mL of kanamycin was inoculated for both BH and CAH and incubated at 37°C until OD<sub>600</sub> of 0.5-0.7 and induced with 1 mM IPTG for 4 hours at 37°C, 250 rpm. The expression and purification protocol is the same as the protocol used for the Src kinase.

**(SI 2.7) Enzyme-linked immunosorbent assay (ELISA)** – Phosphorylated AtzAM1 (pY-AtzAM1) was loaded onto clear flat-bottom immuno 96-well plates (Thermo Scientific item # 442404) at 20µg/mL and 1.25µg/mL in 50µL 1X PBS (Gibco pH 7.4, #10010023) overnight at 4°C. Plates were rinsed twice in 200µL 1X TBS (Biorad

#1706435). 1% BSA in TBS 0.05% Tween 20 was used to block wells at 200 $\mu$ L block solution for 1.5hr at 25°C under gentle agitation. Anti-phosphotyrosine 4G10 Platinum HRP conjugate (EMD #16-316) was diluted 1:5000 in 1% BSA TBS 0.05% Tween 20 and loaded onto the well at 25°C for 1.5hr under gentle agitation. Excess anti-phosphotyrosine was washed off with 200 $\mu$ L of TBS 0.05% Tween 20 in triplicate. To detect bound antibody, 100 $\mu$ L of TMB substrate reagent (Biolegend #421101) was added to each well and incubated for 5 minutes at 25°C. 100 $\mu$ L of TMB stop solution (Biolegend #423001) was added to the wells. Absorbance was read at 450nm using the Tecan Infinite M200 Pro plate reader.

**(SI 2.8) Bio-layer interferometry (BLI)** – AtzAM1 was phosphorylated using the conditions described below. pY-AtzAM1 was then biotinylated at 10mM Sulfo-NHS-Biotin (APExBIO) for 30min at 25°C. Excess biotin was buffer exchanged with a PD-10 desalting column (GE Healthcare) equilibrated with HNG. Biotinylated pY-AtzAM1 was loaded onto streptavidin (SA) coated biosensors (ForteBio) and used for BLI. AtzCM1 was flowed in from 4nM to 4 $\mu$ M. BLI experiments were performed using the BLItz System (ForteBio).

**(SI 2.9) Phosphorylation, assembly formation, and disassembly** – The phosphorylation protocol was based upon Src kinase activity assay by Sigma (Catalog # S1076). In a final reaction volume of 150 $\mu$ L, 3 $\mu$ M AtzAM1 was mixed into 1X Kinase Activity Buffer (4mM MgCl<sub>2</sub>, 2.5mM MnCl<sub>2</sub>, 0.25mM DTT, 5mM MOPS, 2.5mM glycerol-2-phosphate, 1mM EGTA, 400nM EDTA, pH 7.6), 2.5 mM MnCl<sub>2</sub>, HNG, 2 mM ATP, 800ng Src kinase, and incubated for 7 – 16 hr at 25°C for phosphorylation to occur.

After phosphorylating, AtzCM1 was added to a final 2 $\mu$ M concentration. Assembly was allowed to form at 2hr 25°C. Disassembly was performed by adding 4.8 $\mu$ g of YopH phosphatase into the 150 $\mu$ L reaction mixture after assembly formation occurred. Size measurements using DLS were performed to determine assembly formation/disassembly.

**(SI 2.10) Dynamic light scattering (DLS)** – 50  $\mu$ L of an assembly sample was used for size determination using a Malvern Zetasizer and a quartz cuvette (ZEN2112, Malvern). Ten spectra measures were recorded for eleven replicates at 25 °C. The standard operating procedure accounted for 5% glycerol in solution.

**(SI 2.11) DLS Inhibition Experiment** - 6  $\mu$ M pY-AtzAM1 was phosphorylated (1X KAB, 2 mM ATP, 1 mM DTT, HNG, 1  $\mu$ g Src kinase) in a reaction volume of 75  $\mu$ L. Incubation time was overnight at 25°C. SH2 or SH2-DhaA was added to each sample at 0  $\mu$ M, 3  $\mu$ M, 6  $\mu$ M, 9  $\mu$ M, 12  $\mu$ M, 15  $\mu$ M, 18  $\mu$ M final concentration and allowed to "block" binding sites on the pY-AtzAM1 for 1 hr at 25°C. AtzCM1 was added to each sample at 2  $\mu$ M final concentration. Therefore, the final concentrations of all components was 3  $\mu$ M pyAtzA, 1  $\mu$ M AtzCM1, 0  $\mu$ M - 18  $\mu$ M SH2 or SH2-DhaA. The sample was incubated for 2 hr at 25°C. DLS was performed to analyze assembly sizes. DLS was performed at 25°C, 50  $\mu$ L/sample volume, in a low-volume quartz sizing cuvette (Malvern; ZEN2112) using a Zetasizer Nano ZS (Malvern). Measurements were performed in triplicates while each sample was read and averaged 15 times. This protocol was repeated at a final concentration of 1  $\mu$ M pyAtzA, 0.66  $\mu$ M AtzCM1, 0  $\mu$ M -6  $\mu$ M

SH2-DhaA. Curve fitting was performed in MATLAB (R2016b; Mathworks) using the general model:

$$f(x) = \frac{A}{1 + e^{-k*(x-x_0)}} + B$$

where  $A$ ,  $B$ ,  $k$ ,  $x_0$  are constants. Adjusted  $R^2$  was used to determine model validity.

Inhibition concentration 50 (IC50) was determined based upon concentration of inhibitor that resulted in assembly size of 100nm measured.

**(SI 2.12) DLS Titration Experiment** – 6  $\mu$ M, 3  $\mu$ M, 1.5  $\mu$ M, 0.5  $\mu$ M, 0.1  $\mu$ M pyAtzA was phosphorylated (as described previously) with an incubation time of overnight at 25°C. Either AtzCM1 wildtype (WT) or AtzCM1 superbinder (SB) was added to each sample at 2  $\mu$ M, 1  $\mu$ M, 0.5  $\mu$ M, 0.25  $\mu$ M, 0.50  $\mu$ M final concentration. The sample was allowed to incubate for 2 hr at 25°C. Therefore, the final concentrations of all components was from 3  $\mu$ M – 0.05  $\mu$ M pyAtzA, 2  $\mu$ M – 0.05  $\mu$ M AtzCM1-WT or AtzCM1-SB. DLS was performed at 25°C, 50  $\mu$ L/sample volume, in a low-volume quartz sizing cuvette (Malvern; ZEN2112) using a Zetasizer Nano ZS (Malvern). Measurements were performed in duplicate with each sample read and averaged 15 times.

**(SI 2.13) DLS Kinetics (varying ATP) Experiment** – An assembly mixture of 3  $\mu$ M non-pyAtzA and 2  $\mu$ M AtzCM1 was prepared (as described previously) and syringe-filtered at 0.22  $\mu$ m. To each 50  $\mu$ L reaction volume, 1.2  $\mu$ g of src kinase was added. Size was monitored continuously for 30 min at 25°C in a low-volume quartz sizing cuvette

(Malvern; ZEN2112) using a Zetasizer Nano ZS (Malvern) at 50  $\mu\text{L}$ /sample.

Measurements were performed in triplicates. Each sample was read and averaged five times over the course of 25 seconds for a single time point. Curve fitting was performed in MATLAB (R2016b; Mathworks) using sloping spline function, with varying smoothing parameters. Adjusted  $R^2$  was used to determine model validity.

### 8.7.2 SI 3. Microscopy methods

**(SI 3.1) Transmission electron microscope (TEM)** – Assembly (3  $\mu\text{M}$  pY-AtzAM1 and 2  $\mu\text{M}$  AtzCM1) and non-assembly (3  $\mu\text{M}$  non-pyAtzA and 2  $\mu\text{M}$  AtzCM1) samples were mixed, and diluted ten-fold in deionized water. The diluted samples were applied to the carbon-coated FCF400-Cu grids (Electron Microscopy Sciences, Hatfield, PA) which had been glow-discharged for two hours under UV light to render the grids hydrophilic and adsorptive. A drop of sample ( $\sim 5\mu\text{L}$ ) was added on a piece of wax film and the grid was placed onto the sample droplet for absorption for two minutes. Excess sample solution was removed with a filter paper. A drop ( $\sim 5\mu\text{L}$ ) of 1% uranyl acetate was dropped on the wax paper and the grid was placed onto the staining solution droplet for two minutes to stain. Excess staining solution was removed by blotting with a filter paper, the grids were allowed to air dry for two minutes. Images were collected on JEOL 1200EX electron microscope with AMT-XR41 digital camera.

**(SI 3.2) Atomic force microscopy (AFM)** – The assemblies were directly visualized by non-contact mode atomic force microscopy (AFM) Parks Systems. Samples were prepared by depositing 20  $\mu\text{L}$ s of sample on silicon wafer and incubated for 5 minutes. After incubation, the silicon was washed with deionized water to remove salt and air

dried overnight at 25°C. Assemblies were visualized by an AFM (Parks System). The AFM was used in non-contact mode (330 kHz resonant frequency and 42 N/m spring constant, PPP-NCHR Park Systems, #610-1051). Images were taken with 2048x2048 pixels with scan rates of 2  $\mu\text{m/s}$  to 30  $\mu\text{m/s}$ . The AFM images analysis was performed using Gwyddion software<sup>19</sup>.

**(SI 3.3) Helium ion microscopy (HIM)** – The AFM sample preparation on a silicon wafer was used for HIM. Imaging was done on the Carl Zeiss Orion Plus Helium Ion Microscope (Carl Zeiss Microscopy, Peabody, MA) operating at 30 KeV acceleration voltage with a beam currents of about 1 pA. Most samples did not exhibit significant charging therefore electron flood gun was not used for charge neutralization. The vacuum reading in the analysis chamber during imaging was  $2 \times 10^{-7}$  torr.

**(SI 3.4) High-resolution fluorescence microscopy** – For the growth video, 20  $\mu\text{L}$  of 3  $\mu\text{M}$  AtzAM1 and 2  $\mu\text{M}$  AtzCM1 sample (with all the required buffers as described previously) was deposited on a glass cover and 0.2  $\mu\text{m}$  of Src kinase was added to the sample to allow for assembly formation to occur. The sample was monitored for an hour. For the 3-component assembly image (3  $\mu\text{M}$  pY-AtzAM1, 1  $\mu\text{M}$  AtzBSH2, 2  $\mu\text{M}$  AtzCM1) the AtzBSH2 protein was dye labeled with the Alexa Fluor<sup>TM</sup> 647 NHS Ester (Succinimidyl Ester, ThermoFisher Scientific #A2006) and buffer exchanged into HNG with a PD10-desalting column. Fluorescent images along with bright-field images were collected. Images were captured using a Nikon Ti-E inverted microscope. A Coherent Genesis laser at 567 and Coherent Obis Laser at 647 were used for fluorescent imaging, using 1mW power. Images for the assemblies with antibody fluorescence was taken using

a 2048x2048 pixel resolution. A 561 nm laser at 12 mW was used and imaged with a 100x TIRF high NA (1.49) oil immersion objective. Samples were placed on a bacto (tm) agar pad at 1.5% w/v (150 mg/10mL). The agar pad was hardened in a gene frame on a 25 mm coverslip and sealed with an 18 mm coverslip on top.

**(SI 3.5) Cryo-EM Tomographic tilt series acquisition and reconstruction** – For cryo-electron tomography, an AtzAM1 and AtzCM1 assembly sample was mixed with 10 nm gold fiducial markers to facilitate alignment in data processing. An aliquot of 3.5ml sample was applied to 2.0/1.0mm Quantifoil holey grids (Quantifoil, Germany) and plunge frozen using a Leica EM GP plunger (Leica). Tomographic tilt series acquisition was performed on a Talos Arctica microscope (Thermo Fisher) operated at an acceleration voltage of 200kV. This microscope was equipped with a field-emission gun, Volt phase plates, Gatan postcolumn energy filter and a K2 summit direct electron detector. Tilt series were collected at 39,000x microscope magnification with -0.5  $\mu\text{m}$  defocus using FEI Tomography software. The sampling of the data was calibrated to be 3.49  $\text{\AA}$ /pixel. Typically, a tilt series ranged from -60° to 60° at 3° step increment. The accumulated dose for each tilt series was 60 electrons/ $\text{\AA}^2$ . Tilt series were aligned based on fiducial gold markers using the IMOD package<sup>20</sup>. 3D tomograms were obtained by weighted backprojection of aligned tilt series. Visualization and annotation of the 3D volumes were done in Chimera<sup>21</sup>.

**(SI 3.6) Cryo-EM AtzAM1 and AtzCM1 model fitting and statistical analysis** – AtzAM1 and AtzCM1 complex subtomograms were extracted from 3D tomograms and bandpass filtered to reduce high frequency noises and low frequency gradient from ice thickness variation. Centers of AtzAM1 and AtzCM1 densities were identified as peaks

within solid voxel clusters that were approximately sizes of an AtzAM1 hexamer, or an AtzCM1 tetramer. Potential free AtzAM1 or AtzCM1 complexes that were too close to a neighboring voxel peak ( $<120\text{\AA}$ ) were removed. Assignment of AtzAM1 or AtzCM1 to an identified voxel cluster was done by applying the condition that AtzAM1 and AtzCM1 alternate in a chain. Densities that had three or more linkers to neighbors were assigned to be AtzAM1. Linear, unbranched assemblies were assigned by first determining identity of one end based on cross-correlation scores between the end peak densities and AtzAM1 or AtzCM1 models computed from their PDB structures. Assignment conflicts were resolved by pruning along the branches in the order of intensity values. The above protocol was first applied to a small assembly, and optimized and validated by human visual inspection before it was used on larger assemblies. Coordinates and connection information of each AtzAM1 or AtzCM1 complex in an assembly were extracted and used for statistical analysis and for comparison to simulation data. The volume of the assembly is defined by the volume of the convex hull that encloses all determined AtzAM1 or AtzCM1 molecule.

**(SI 3.7) Confocal microscopy fluorescent images of fractal and globular assembly**

**with GFP-SH2 and Goat anti-mouse IgG Antibody** - Fluorescently tagged samples were placed in chamber slides and allowed to air dry overnight. Fluorescent images were acquired using a spinning disc confocal microscope (Olympus DSU-IX81) fitted with 482nm and 543nm excitation filters and emission filters of 536nm and 593nm, respectively. Sample images were obtained using the 3D image capture function (Z-stacks) with an approximate depth of  $200\mu\text{m}$  at  $1\mu\text{m}$  intervals (step size) using an oil immersion objective (Olympus UPlanFL N 40X/1.3 Oil) and 300ms as exposure time.

Image processing was performed with SlideBook 5.0 (3i, Intelligent Imaging Innovations).

### 8.7.3 SI 4 Enzymatic and Fractal-incorporation Assays

**(SI 4.1) Enzymatic activity measured using the Berthelot assay** – Assembled enzyme samples (1.5  $\mu$ M AtzAM1, 0.5  $\mu$ M AtzBSH2, and 1  $\mu$ M AtzCM1) were made by incubating the enzymes in 1X kinase activity buffer (with no DTT), 2.5 mM MnCl<sub>2</sub>, HNG, 0.2  $\mu$ M Src kinase, and 2 mM ATP in a total volume of 500  $\mu$ l at 25°C for 4 hours. The unassembled enzyme samples were prepared using the same conditions, except no ATP was added to the sample. DLS was performed to verify assembly formation. 10  $\mu$ L of 20 mM Atrazine dissolved in methanol was added to each 500  $\mu$ L sample, for a final concentration of 400  $\mu$ M atrazine, and another sample with the same conditions had no substrate added in order to establish a baseline measurement. Each condition was done in triplicate. After the addition of substrate, the samples are shaken at 100 RPM for 1.5 hr at 25°C. 140  $\mu$ L of each sample is transferred to PCR tubes, then boiled at 99°C for 1.5 minutes, and then cooled at 4°C. The 140  $\mu$ l were transferred to 1.5 mL microcentrifuge tubes and spun down at 20,000 rcf for 20 minutes to remove precipitated protein. 80  $\mu$ l of the supernatant was used for the following steps. 1 $\mu$ g per 20  $\mu$ L of sample of CAH and 1 $\mu$ g per 20  $\mu$ L of sample of BH was added to each sample. The samples were incubated at 25°C for 2 hours to allow for the complete conversion of the cyanuric acid to ammonia by CAH and BH. The Berthelot assay was performed in triplicate on the resulting samples to determine the production of ammonia. For every mole of cyanuric acid produced, one mole of ammonia was assumed to have been

produced. 20  $\mu\text{L}$  of each sample was added to a 96-well plate (Greiner half area clear #675101). 60  $\mu\text{L}$  of solution A (0.05 g/L sodium nitroprusside and 10g/L phenol) was added and mixed into every sample. Then 80  $\mu\text{L}$  of solution B (5 g/L NaOH and 8.4 mL/L bleach) was added and mixed into every sample. The samples were incubated for 30 minutes at 25°C for a blue color to develop. The absorbance at 630 nm was read using Tecan Infinite M200 Pro plate reader. The extinction coefficient was determined using standards of cyanuric acid at known concentrations in the enzyme activity buffer that had been reacted with the BH and CAH for 2 hours.

**(SI 4.2) Temperature stress activity assays** – Assembled and unassembled enzyme samples were made as described above and incubated at 25°C for 4 hours to allow full assembly formation. The assemblies were then incubated at the following temperatures: 25°C, 40°C, 45°C, 50°C, 55°C, and 60°C for fifteen minutes, and cooled back to 25°C before the addition of 400  $\mu\text{M}$  atrazine. After atrazine was added, the enzyme activity assay was performed as described above.

**(SI 4.3) Shaking stress activity assay** – Assembled and unassembled enzyme samples were made as described above and incubated at 25°C 4 hours. Both samples were shaken at 50, 100, 150, 200, 225, and 250 RPM 25°C for 1 hour before any addition of atrazine. 400  $\mu\text{M}$  atrazine was added to the samples and shaking continued at their respective shaking speeds for 1.5 hour. The rest of the activity assay protocol was conducted the same as described above.

**(SI 4.4) Construction and assay of Basotect® polymer foam with trapped assemblies and free enzymes** – Hydrolyzed TEOS was prepared by combining 7 ml TEOS (Aldrich #131903), 3 ml water, and 0.04 ml 0.1N hydrochloric acid and stirring the solution for 2

hr at room temperature<sup>22</sup>. Basotect® polymer foam (Procter and Gamble UPC# 0 37000 43515 0) was cut into 2.0 x 2.0 x 0.3 cm squares with a razor and 0.250 ml of assemblies or free enzyme solution was spotted onto each 2 x 2 cm face of the foam squares. Aliquots (1.0 or 0.5 ml) of hydrolyzed TEOS were diluted with HNG buffer to a final volume of 10 ml (10% or 5% TEOS). A single application of 5% or 10% hydrolyzed TEOS solutions was done with a small paint brush (Richeson 95822). The TEOS was allowed to set for 2 h, and then liquid was squeezed out of each foam square and total protein concentration in the liquid was measured with the Bradford assay (BioRad #500-0006). To assay activity in the embedded foam, 1 ml of 150  $\mu$ M atrazine in 1X phosphate buffered saline (pH 7.4) was soaked into the foam squares and incubated for 1.5 hour at 25°C. Liquid was squeezed out after incubation and boiled as above to inactivate eluted enzymes. Cyanuric acid produced during the incubation was assayed as described except that the Berthelot reactions were conducted in 10 x 4 x 45 mm cuvettes (Sarstedt #67-742) and read using a Beckman DU 640 spectrophotometer.

**(SI 4.5) GFP-SH2 incorporation assays** – AtzAM1 and AtzCM1 (along with AtzA/AtzC extended linker versions for globular assemblies) assemblies were formed in a reaction volume of 3 mL, 15  $\mu$ M AtzAM1 and 10  $\mu$ M AtzCM1 into 1X Kinase Activity Buffer, 2.5 mM  $\text{MnCl}_2$ , HNG, 2 mM ATP, 0.2  $\mu$ M Src Kinase, and allowed to form for 10 minutes before the addition of 1.8  $\mu$ M Gfp-Sh2 protein, and incubated for 4 hr at 25°C for phosphorylation to occur. Samples were spun down for 2 minutes (500 x g) and supernatant measured in a black half-area microplate (excitation 395 nm, emission 509 nm) with a gain of 140 on a Tecan Infinite M200 Pro plate reader.

**(SI 4.6) DhaA-SH2 incorporation assays** – AtzAM1 and AtzCM1 (along with AtzA/AtzC extended linker versions for globular assemblies) assemblies were formed in a reaction volume of 3 mL, 15  $\mu$ M AtzAM1 and 10  $\mu$ M AtzCM1 into 1X Kinase Activity Buffer, 2.5 mM  $\text{MnCl}_2$ , HNG, 2 mM ATP, 0.2  $\mu$ M Src Kinase, and allowed to form for 10 minutes before the addition of 1.8  $\mu$ M of DhaA-Sh2, and incubated for 4 hr at 25°C. Assemblies were spun down and pellet resuspended with 10 mM TCP and incubated for 1 and 16 hr. Assemblies were spun down again and supernatant was measured at A560 nm.

**(SI 4.7) Goat anti-mouse IgG incorporation assays** - AtzAM1 and AtzCM1 (along with AtzA/AtzC extended linker versions for globular assemblies) assemblies were formed in a reaction volume of 1.8 mL, 12  $\mu$ M AtzAM1 and 8  $\mu$ M AtzCM1 into 1X Kinase Activity Buffer, 2.5 mM  $\text{MnCl}_2$ , HNG, 2 mM ATP, 0.2  $\mu$ M Src Kinase, and allowed to form for 10 minutes before the addition of 8  $\mu$ M of ProteinA-Sh2, and incubated for 4 hr at 25°C. Assemblies were spun down 20,000 x g for 20 min in order to measure fluorescence in supernatant. For disassembly assays with YopH, assembly pellets were spun down 20,000 x g for 20 min, supernatant removed, pellets washed with HNG buffer, and spun down again to remove wash, and resuspended in HNG containing YopH. Assemblies were left shaking at 100 RPM for 12 and 24 hrs at 25°C. Assemblies were spun down again and supernatant measured for released antibody.

#### 8.7.4 Supplementary Discussion

#### **(SI. 5.1) Molecular features determining fractal formation**

It has been demonstrated<sup>23</sup> that atomic-level control over component placement is necessary to achieve via computational design, periodic, regularly ordered 2D protein lattices<sup>24</sup> or closed form 3D icosahedra<sup>25</sup>. In contrast, where 2D lattices and 3D closed form assemblies require exacting orientation and rigidity of inter-protein components, fractal assemblies require a degree of flexibility leading to degeneracy of binding modes and anisotropy at the interface of protein components. However, the amount of flexibility needs to be tuned: too little and crystal lattices will form (Fig. S1), too much and globular protein agglomerates will result (as observed for our control assemblies involving long loop connectors).

To obtain a fractal assembly with protein components, based on our data, we hypothesize that three tunable factors contribute: valency, affinity, and flexibility.

Valency, the measure of possible favorable connections between protein components, contributes to the amount of branching as well as the orientation of the inter-protein components. With homomeric D<sub>2</sub> and D<sub>3</sub> protein components, we anticipated the D<sub>3</sub> (AtzA) to make up to 6 connections to the D<sub>2</sub> (AtzC) component which is capable of 4 connections. If the affinity of the inter-protein connection is sufficiently strong, and the length of the bridging interactions is kept short we could observe an avidity effect between components—where two bridges (divalent connection) are formed between two components (Fig. 1). The formation of divalent bridging connections, localized to C<sub>2</sub> sub-symmetries of D-symmetric proteins, can greatly reduce the flexibility between connected protein components while providing high affinity. In this way, avidity and symmetry together can be utilized to introduce orientational anisotropy and rigidity of the

inter-component connections.

To promote avidity, we chose strong (nM affinity) peptide-binding motifs which could be fused to the D-symmetric protein building blocks. During design, to ensure that any divalent connections made between components were restricted to connections along the C<sub>2</sub> sub-symmetry axes, we imposed design constraints on the fusion linker lengths—maintaining that no additional residues would be added beyond the residues found in the crystallographic structure files creating a direct fusion (0-residue linker) and introducing Gly and Ser residues (GGS) to promote (limited) flexibility.

#### **(SI. 5.2) Fractal dimension and lacunarity**

The fractal (Hausdorff-Besicovitch or box counting) dimension<sup>26</sup>, a general measure of how the size of an object scales as a function of the size of its building blocks, has been used to characterize simulated/mathematical fractal patterns<sup>27,28</sup> as well as real-world statistical fractals including peptide-based fractals obtained on a surface and imaged with AFM<sup>29</sup>. Intuitively, the dimension of an object can be thought of as the scaling observed for change in its overall size (or mass) upon changing the size of its unit building block. For example, a square has dimension two because its mass (proportional to area) grows with a scaling exponent of 2 with the length of its side (if length is increased  $n$ -fold then area increases by  $n^2$ -fold). A cube has a dimension of three because its mass (volume) will increase by  $n^3$  times when length of its side is increased  $n$ -fold. For some objects, this scaling is observed to non-integer and captures how the object occupies space. For example, a curve with a fractal dimension 2.1 fills space very much like an ordinary 2-dimensional surface, but a curve with a fractal dimension of 2.9 folds to fill space nearly

like a 3-dimensional volume. However, any arbitrary curve is not fractal – it has to follow a scaling equation throughout the space it occupies as described below. Another intuitive heuristic is that fractal dimension can be thought of as a measure of the "roughness" of an object's periphery ("Clouds are not spheres, lightening is not a straight line" – Mandelbrot). The surface of the human brain, for example, is fractal.

To calculate fractal dimension for a fractal  $S$ , we consider the object  $S$  lying on an evenly spaced grid (of size  $L_0$ ), and count how many boxes are filled by elements of  $S$ . The box-counting dimension is calculated by seeing how this number changes as we make the grid size ( $L$ ) finer by applying a box-counting algorithm as shown in Fig. S4 (using an image obtained from microscopy in our study). Thus, the fractal dimension,  $D_f$ ,<sup>30–34</sup> calculated from box counting is obtained:

$$\text{Log}(N) = D_f \cdot \text{Log}\left(\frac{L}{L_0}\right)$$

The existence of a straight line with a single slope in the above plot indicates that the object is fractal and indicates statistical self-similarity. If different slopes are obtained at different scales, the object is considered multi-fractal.

Another quantity that captures the "gappiness" or "holeyness" of a fractal shaped object is lacunarity ( $\lambda$ )<sup>30</sup>. In box counting analyses, lacunarity  $\lambda$  for each grid of calibre  $\varepsilon$  ( $= L/L_0$ ) is calculated from the standard deviation,  $\sigma$ , and mean,  $\mu$ , for pixels per box. That is, there is a  $\lambda$  value for each  $\varepsilon$  in each series of grid sizes in each grid orientation in a set of grid orientations.

$$\lambda_{\varepsilon,g} = CV_{\varepsilon,g^2} = \left( \frac{\sigma_{\varepsilon,g}}{\mu_{\varepsilon,g}} \right)^2$$

This value is averaged over all orientations to obtain the reported lacunarity,  $\lambda$ , value. We used the implementation in ImageJ software to calculate fractal dimension and lacunarity.

We note that in our analyses, fractals formed by the same components can vary in shape and dimension from isolated assembly on the surface (island) to island as well as in solution. However, despite inter-island variations, every island is self-similar (with the same fractal dimension). Similar type of topological diversity (and uniformity within islands) was also found in studies of silk protein sericin<sup>34</sup>, where variation in fractal dimension of observed protein islands was detected depending on the surface conditions but each island was self-similar. For all 2D image analyses in this paper, we derived the fractal dimension (slope), scalability (linear range), and lacunarity from 2D image analysis using ImageJ. Due to the island-to-island variation, all 2D-analyzed  $D_f$  and  $\lambda$  values reported in this work are an average of at least 5 individual islands (as many as 20 images were used when available).

When comparing the Cryo-ET data to the computational simulation results, projections were made to be analyzed with the same 2D image analysis. Additionally, 3D-fractal dimension analysis was performed with an in-house 3D box-counting algorithm that works in the same way that 2D image analysis does except the two-dimensional boxes are replaced with three-dimensional cubes (voxels) during the scaling analysis.

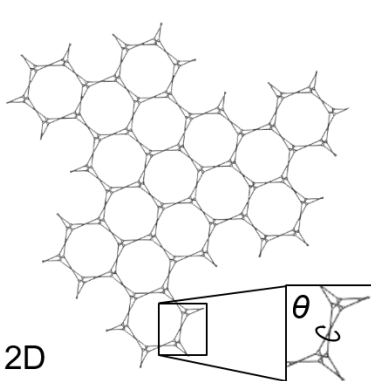
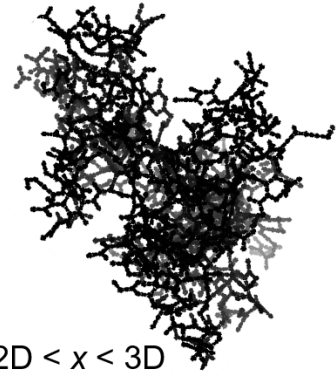
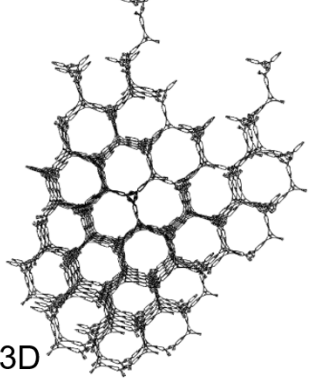
**(SI. 5.3) Comparison of control (GS-rich-linker containing) and designed assembly topologies using cryo-ET**

Although we could resolve the cryo-ET-derived density of the fractal assemblies (Fig. S26A and Fig. S26C, respectively), the globular (GS-rich-linker containing) assemblies varied too greatly in topology across samples to analyze—the majority of these images were dominated by dark shadowy particles too dense to obtain meaningful assignments of density to individual protein components (Fig. S26B). However, a few images (<10%) from the GS-linker rich set had small resolvable nm-scale regions where density could be interpreted and assigned to individual protein components (Fig. S26D). For these images, we compared the average monomer-monomer distance across 5 control (GS-rich) and 5 fractal-shaped assemblies (Fig. S27) on the nm-scale. In the fractal-shaped assemblies, the inter-monomer distance is tightly clustered ( $134 \pm 2 \text{ \AA}$ ) among images of large (>25 nm size) assemblies (~40% of the set), suggesting uniformity of inter-component connections in agreement with the design conception. In contrast, in the resolvable parts of the control assembly tomograms (<10% of the entire imaged sample), we see three different types of structures: dispersed assembly (inter-monomer distance  $\sim 157 \text{ \AA}$ ), fractal-similar assemblies ( $\sim 134 \text{ \AA}$ ), and densely packed globular ball-like structures ( $\sim 125 \text{ \AA}$ ). These data suggest that GS-rich linker conformational flexibility (1) abrogates anisotropy of inter-component interactions and (2) allows reorganization of assembly structure to yield a more globular structure on the micron scale driven by non-specific protein-protein sticking. The robust catalytic activity of the control assembly (Fig. S37) demonstrates that the observed topologies in the control tomograms are not the result of protein unfolding but are in fact, mediated by the SH2 domain-pY peptide interactions.

#### **(SI. 5.4) Evaluating the effects of AtzB-SH2 on overall fractal structure and topology**

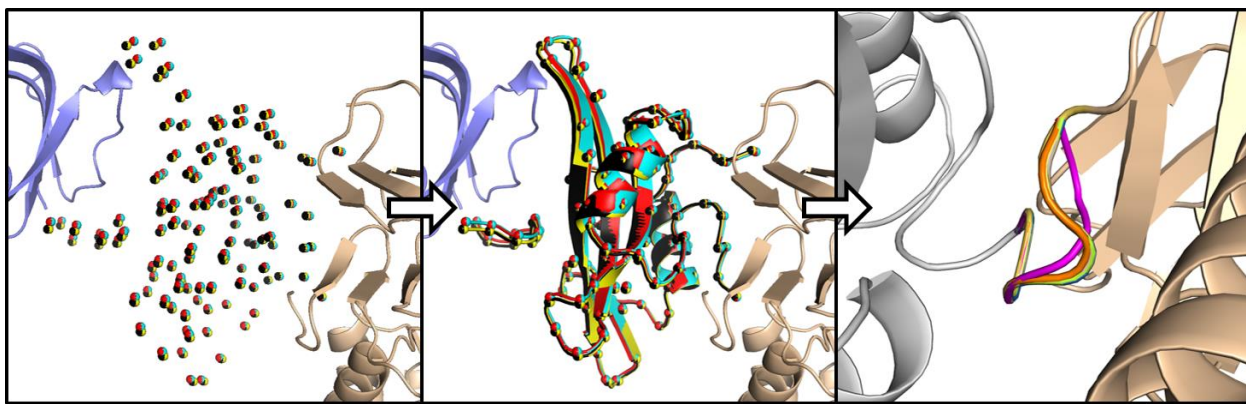
Upon addition of 1 unit AtzB-SH2 to the 3A:2C ratio (added before phosphorylation) we find that the observed fractals show a marked decrease in both  $D_f$  and  $\lambda$  compared to the 3A:2C fractals (Fig.S32). When the concentration of A is increased in the three-component assembly we once again see a decrease in  $D_f$ ; however, we also see a decrease in  $\lambda$  (Table S3, Fig. S31, S32). The observed  $D_f$  and  $\lambda$  in the three-component assembly resemble the  $D_f$  and  $\lambda$  values in the two-component fractals with low concentrations of A relative to C (Fig. S21, Table S3 ). These findings, combined with the observed incorporation of dye-labeled AtzB-SH2 data (Fig. S34 and Fig. 35D), suggest that AtzB-SH2 is competing for locations to bind the SH2-peptide fused to AtzAM1 and is further changing the structural features of the assembly.

## 8.7.5 SI Figures and Tables

 <p>2D</p>	 <p><math>2D &lt; x &lt; 3D</math></p>	 <p>3D</p>
<p>Choose 1: <math>\theta \in \{0, 90, 180, 270\}</math></p>	<p>Choose &gt;1: <math>\theta \in \{0, 35.25, 54.75, 90, 125.25, 144.75, 180, 215.25, 234.75, 270, 305.25, 324.75\}</math></p>	<p>Choose 1: <math>\theta \in \{35.25, 54.75, 125.25, 144.75, 215.25, 234.75, 305.25, 324.75\}</math></p>
<p><math>\sim 4,000 \text{ units}/\mu\text{m}^2</math></p>	<p><math>\sim 70,000 \text{ units}/\mu\text{m}^3</math></p>	<p><math>\sim 40,000 \text{ units}/\mu\text{m}^3</math></p>

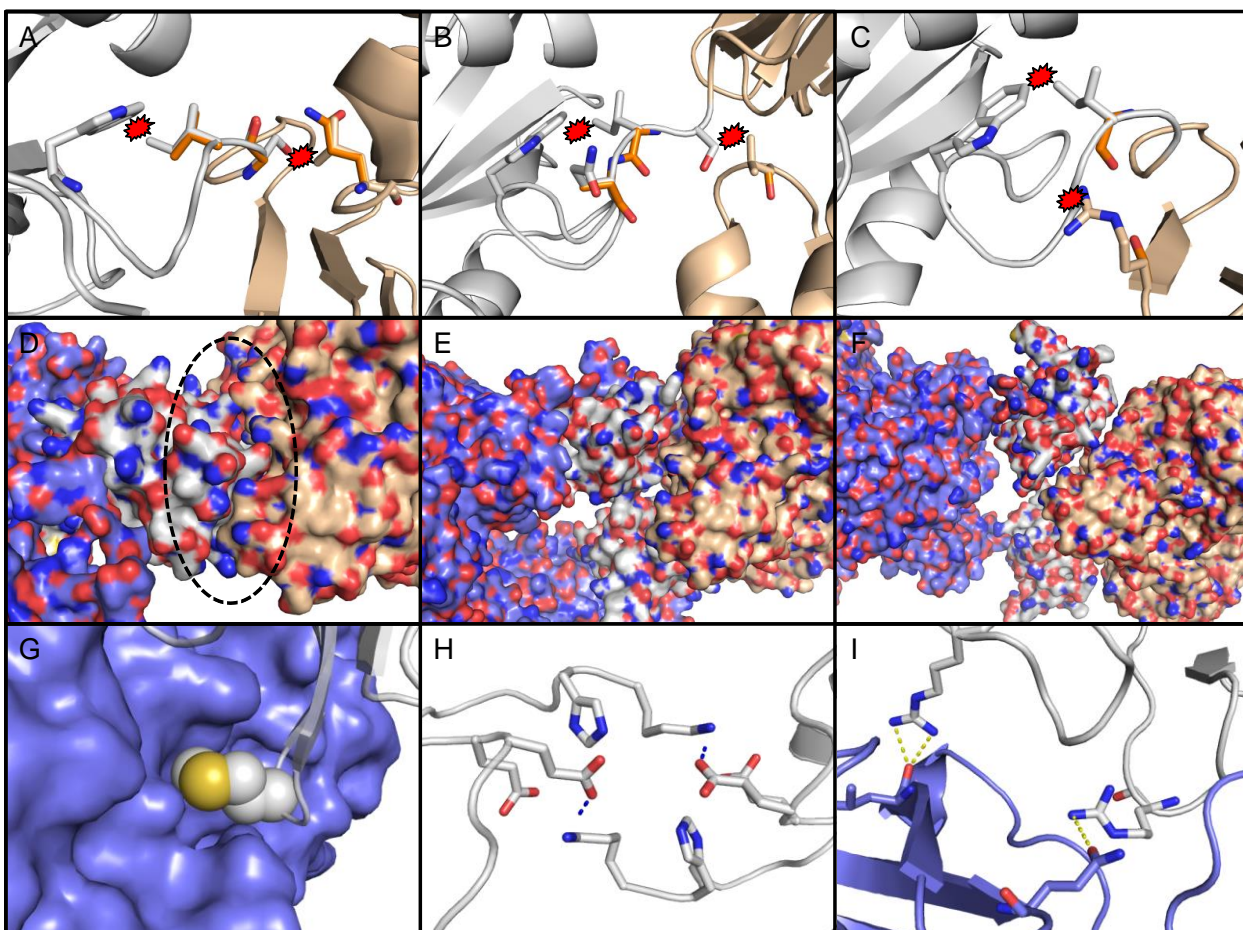
**Figure 96-8.** Supplementary Figure 1. Scheme for designing arboreal fractal morphologies.

Predicted assembly topology based on values of  $\theta$ . For 2D and 3D crystal geometries  $\theta$  would have to be one discrete angle from the angles listed. A fractal would form from stochastic combinations of propagatable  $\theta$  values in the range  $0 \leq \theta < 360$ . For a cube with length  $1\mu\text{m}$  we calculated the number of theoretical total connected protein units that the 2D and 3D assembly built with AtzC-SH2 and AtzA-pY could occupy to be 4000 and 40,000 units respectively and compared that to the number of  $\text{units}/\mu\text{m}^3$  observed for the Cryo-ET characterized fractal assembly ( $70,000 \text{ units}/\mu\text{m}^3$ ).



**Figure 97-8.** Supplementary Figure 2. Flowchart of interface and linker design method.

(A) The SH2 and peptide binding partner alpha carbons are converted into a ligand file which is used by RosettaMatch to determine placements of the domain-peptide pair (colored red, yellow, cyan, and black for 4 unique solutions) within placements of AtzA and AtzCmonomers (with varied  $d$  and  $\theta$ ). B) The ligand is converted to a full atom model by threading (using the alpha carbons) followed by C) fusion loop closure using Kinematic Loop Closure (KIC).



**Figure 98-8.** Supplementary Figure 3. Design considerations for selecting substitutions and atomic interactions responsible for orientations of components during simulation.

Substitutions were introduced if they removed clashes (A-C), supported shape complementarity of novel interfaces (D-F), or created new favorable contacts across the components (G-I). (A-C) For all images, the AtzC monomer is colored in wheat, the SH2 is colored in white, and AtzA monomer colored in slate. Three different design models at unique core rotations: (A) AtzCM2 at 54.75°, (B) AtzCM3 at 54.75°, and (C) AtzCM5 at 0°. Clashes are denoted with a red star and are defined as heavy atoms within 3Å. (D-F) Novel interface complementarity considered in substitution selection. Substitutions that led to favorable shape complementarity and high surface area contact at the novel interfaces (D & E) were kept while design models that showed poor shape complementarity and low surface area contact were rejected (F). Shape complementarity along with novel hydrophobic (G), electrostatic (H), and polar contacts (I) created at novel interfaces are responsible for the decisions made during the stochastic fractal growth simulation.

Table 15-9. Supplementary Table 1. List of substitutions and reasons for the various AtzA and AtzC designs.

	Substitution	Reason	Rotations (within 40 REU of lowest energy)
AtzAM1	I515Y	Polarity	0, 35.25, 54.75, 90, 125.25, 144.75, 180,
	E557S	Clash	
AtzAM2	I516N	Polarity	0, 125.25, 144.75
	Q518G	Clash, flexibility	
	T519P	Clash, rigidity	
AtzAM3	I516N	Polar contact	0, 35.25, 54.75, 90, 125.25, 144.75, 180, 305.25
	Q518G	Clash, flexibility	
	T519G	Clash, flexibility	
AtzAM4	I516D	Electrostatic	0, 35.25, 54.75, 90, 125.25, 144.75, 180, 305.26
	Q518G	Clash, flexibility	
	T519G	Clash, flexibility	
AtzAM5	I516D	Electrostatic	0, 54.75, 90, 125.25, 144.75, 180, 305.27
	Q518G	Clash, flexibility	
	R960H	Clash	
AtzCM1	V402G	Clash, Flexibility	0, 54.75, 90, 125.25, 144.75, 180
	I403G	Clash, Flexibility	
	Q404S	Clash, H-bond to BB	
AtzCM2	L148Q	H-bond to linker BB	0, 35.25, 54.75, 90, 125.25, 144.75, 180, 324.75
	V400G	Flexibility	
	S402G	Clash	
	I403V	Clash	
	Q404A	Clash	
AtzCM3	K40G	Clash	0, 54.75, 90, 125.25, 144.75, 180
	L148S	Clash	
	V400I	Hydrophobic contact	
	I403G	Clash, Flexibility	
	Q404S	Hydrophobic contact	
AtzCM4	R391A	Clash	0, 54.75, 90, 125.25, 144.75, 180
	V398Y	Polar contact	
	V400L	Hydrophobic contact	
	V402G	Clash, Flexibility	
	I403G	Clash, Flexibility	
	Q404S	Clash, H-bond to BB	
AtzCM5	R391S	Clash, H-bond to linker BB	0, 54.75, 90, 125.25, 144.75, 180
	I393C	Clash	
	V398A	Clash	
	V400M	Clash	
	I403G	Clash, Flexibility	
	Q404S	Clash, H-bond to BB	

Lowest energy = -3150 REU, Polarity = surface hydrophobicity reduced, Clash = removed clashing sidechains, Flexibility = introduced backbone flexibility, Rigidity = reduced backbone flexibility, Polar Contact = introduced non-salt bridge side chain hydrogen bonds, Electrostatic = introduced a salt bridge, H-bond to BB = introduced a hydrogen bond to backbone atoms, H-bond to linker BB = introduced a hydrogen bond to the linker backbone atoms, Hydrophobic Contact = introduced a non-polar sidechain interaction

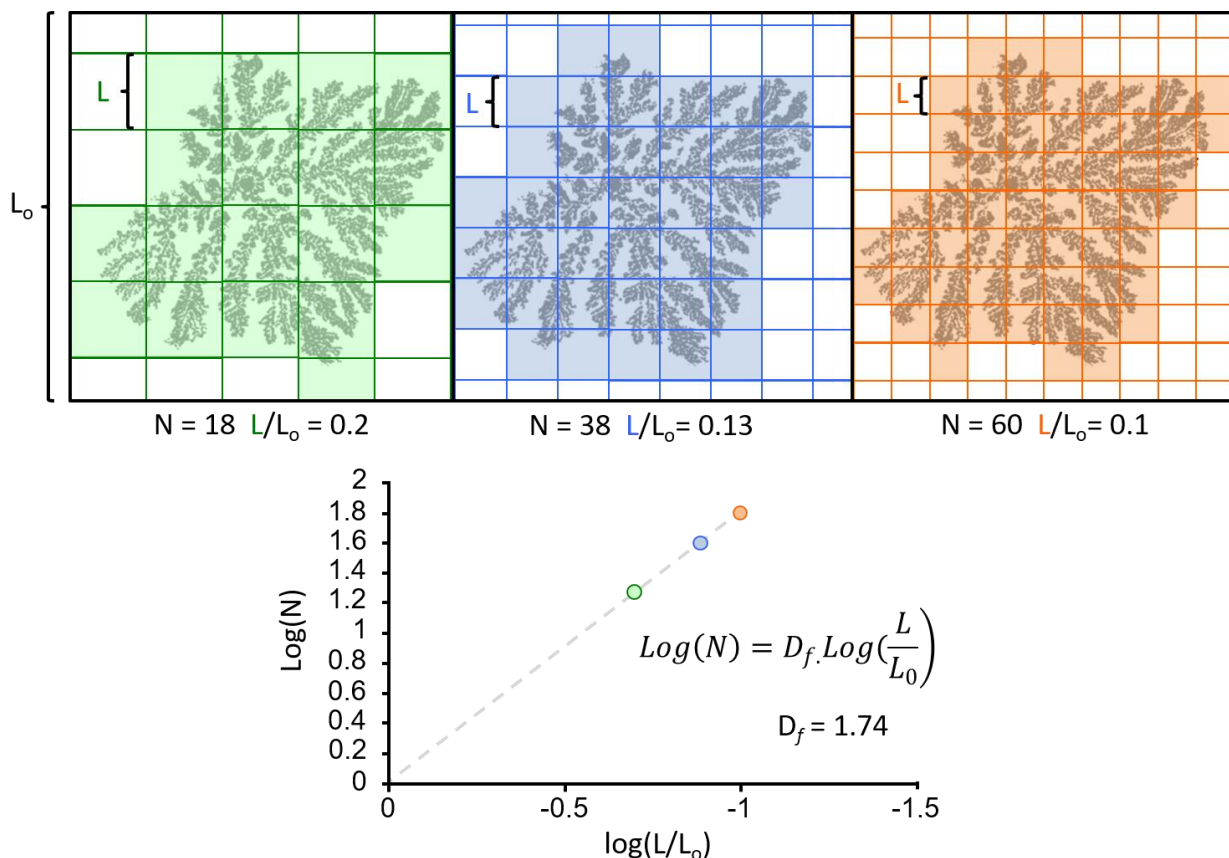
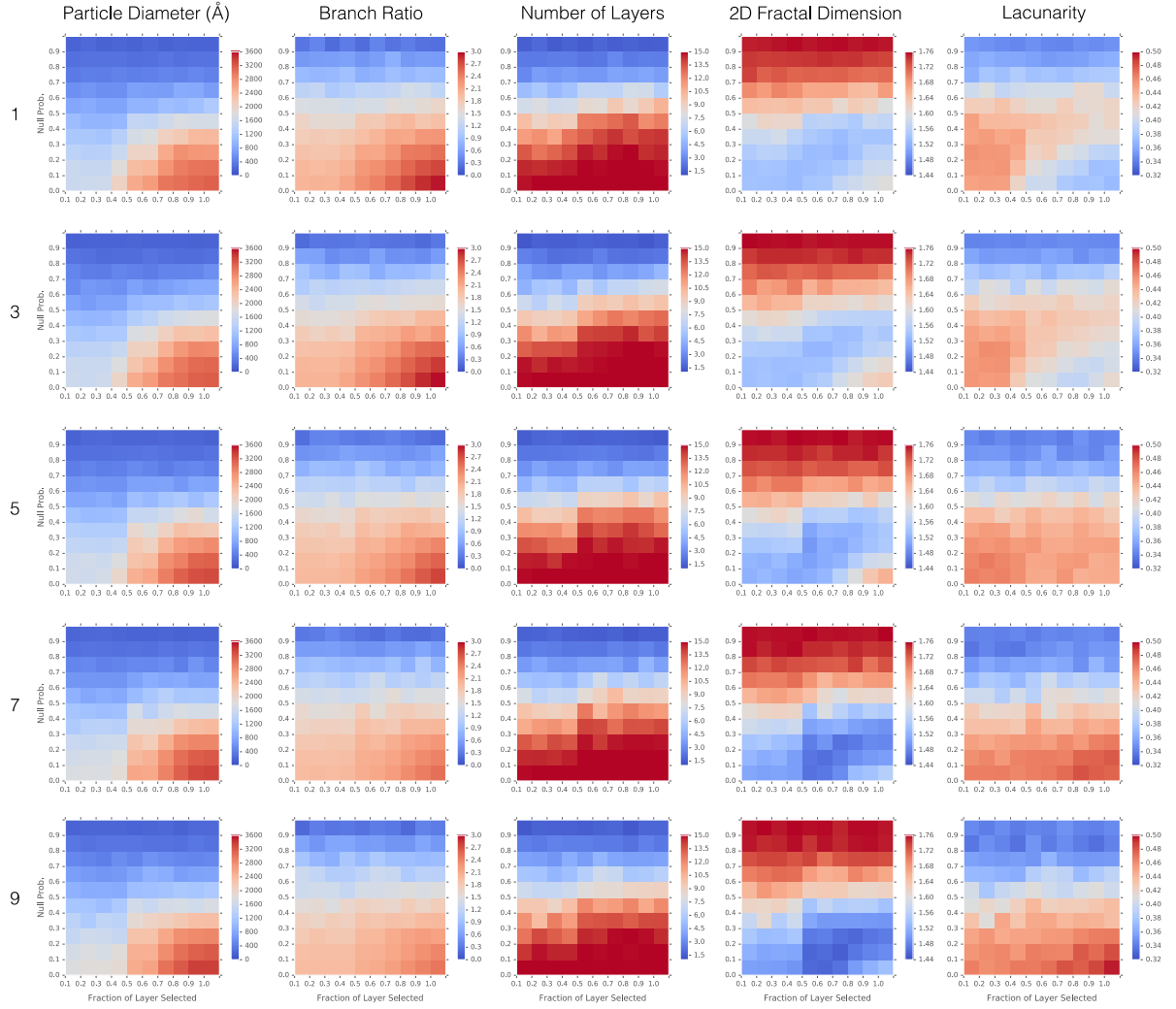


Figure 99-8. Supplementary Figure 4. Illustration of the box-counting algorithm.

Analysis of a fractal image using the box counting method. As box size ( $L$ ) is scaled down from left to right, the ratio of relative box size to the initial box decreases while the number of boxes (shaded boxes represent  $N$ ) that contain the fractal (black image on white background) increases. The log/log relationship (slope) represents the fractal dimension ( $D_f = 1.74$ ). The individual colored points on the graph represent the respective colored image. Coarse features (large  $L$ ) are represented closer to 0 while fine features (small  $L$ ) are more negative. This illustration is just a conceptual representation of the actual method and is not an accurate representation of the fractal dimension. Programs like ImageJ will decrease  $L$  to nearly a pixel in length.



**Figure 100-8.** Supplementary Figure 5. Computational parameter sweep of  $kT$  (major y-axis),  $P_{\text{term}}$  (minor y-axis), and  $C_{\text{frac}}$  (minor x-axis).

The various fractal topologies (limited to 15 layers) were evaluated by their particle diameter, branch ratio, layer count, 2D fractal dimension ( $D_f$ ), and Lacunarity. We observe size, shape, and composition trends with varying  $P_{\text{term}}$  and  $C_{\text{frac}}$ . Less obvious trends in topology via lacunarity and  $D_f$  are also observed with changing  $kT$ .  $P_{\text{term}}$  values above 0.4 (0.5–0.9) and  $C_{\text{frac}}$  values below 0.5 (0.0–0.4) show a steep decline in particle size and number of total layers on average—terminating growth during simulation (unlike experimental data). For non-terminating values of  $P_{\text{term}}$  (0.0–0.4) and  $C_{\text{frac}}$  (0.5–1.0),  $D_f$  is high ( $\sim 1.7$ ) when the connection probability is high—more isotropic fractal—and low ( $\sim 1.6$ ) when the connection probability is low—more anisotropic fractal shapes. When the  $kT$  increases we notice that the relative difference between high and low connection probability is maintained, however, the overall  $D_f$  decreases ( $\sim 1.6$  and  $\sim 1.5$ ) respectively. This can be attributed to the flatter probability landscape allowing for more  $180^\circ$  bound-angle (mixed vertex

and edge centered connections around AtzA)—linearizing the branch connections on average and subsequently decreasing the fractal dimension.

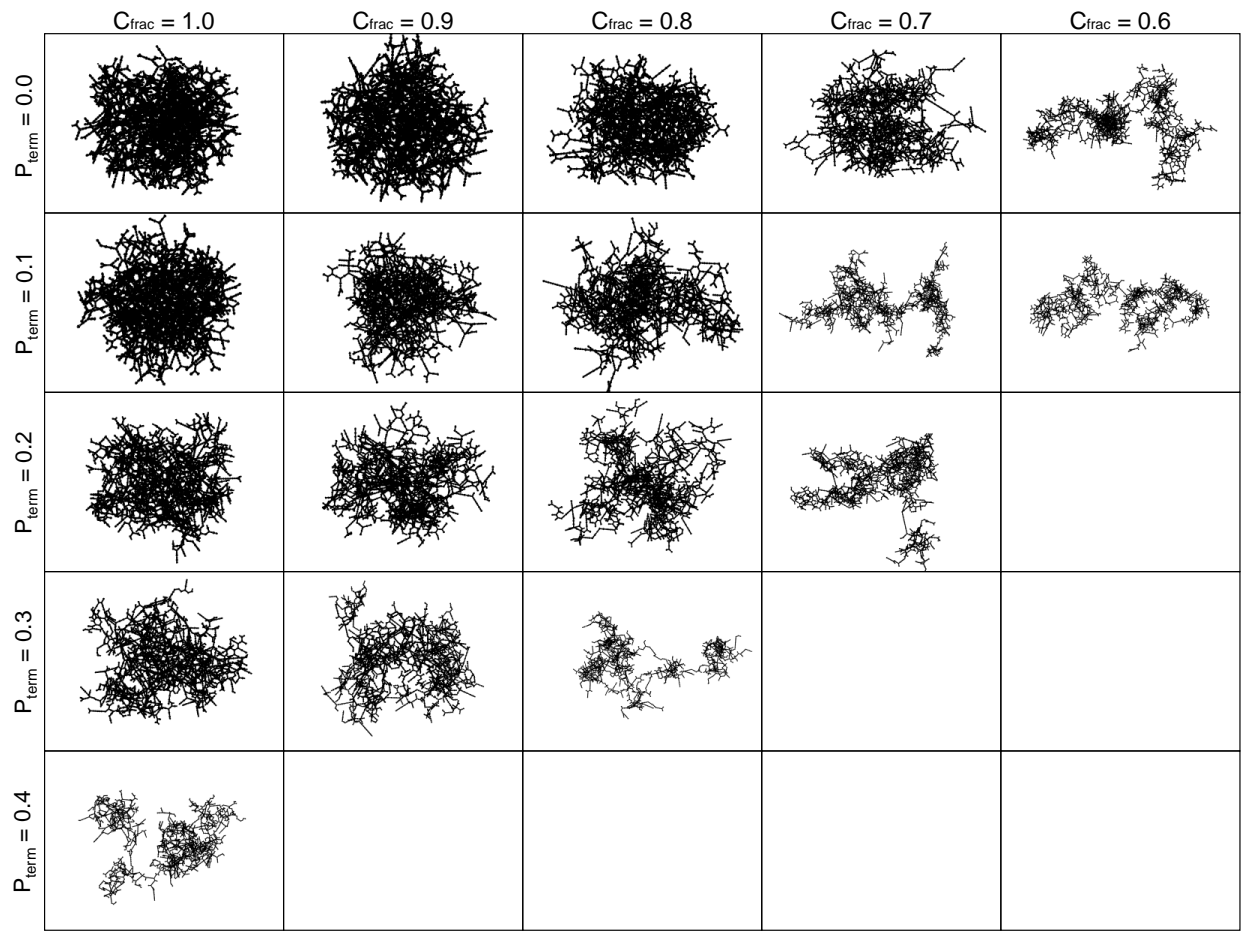


Figure 101-8. Supplementary Figure 6. Representative simulated fractal images

(approx. 5000 components each and  $kT = 9$ ) that possess the average layer count and branch ratio for varying values of  $P_{\text{term}}$  (y-axis) and  $C_{\text{frac}}$  (x-axis) of 100 models. As  $C_{\text{frac}}$  decreases (or  $P_{\text{term}}$  increases) fractal dimension decreases and lacunarity increases.

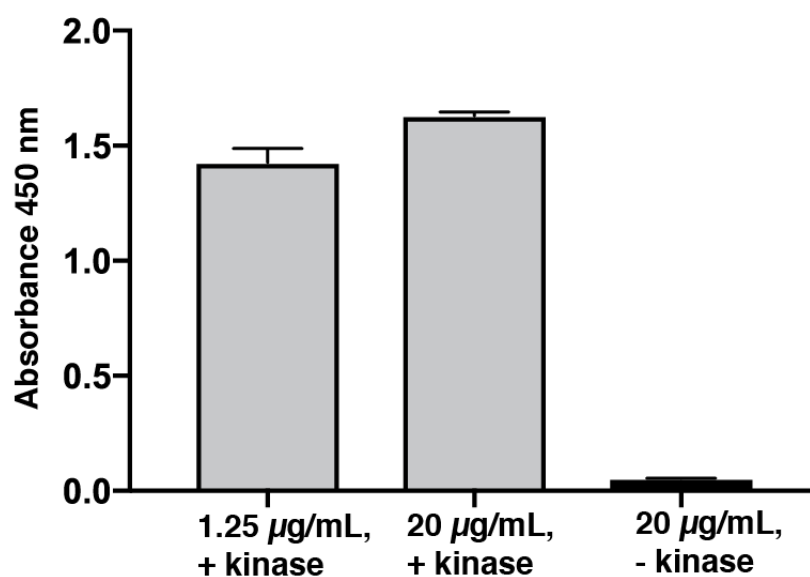


Figure 102-8. Supplementary Figure 7. Phosphorylation of SH2 peptide AtzA fusion (pY-AtzA) by Src kinase.

In order to verify phosphorylation of AtzA by Src kinase into phosphorylated SH2 peptide AtzA fusion (pY-AtzA), ELISA with (1:4000 dilution) antiphosphotyrosine-horseradish peroxidase conjugate was performed on pY-AtzA samples either with Src kinase (+) or without Src kinase (-), in phosphorylation reaction buffer at 1.25 µg/mL pY-AtzA or 20 µg/mL pY-AtzA. Data is presented as mean  $\pm$  1 standard deviation.

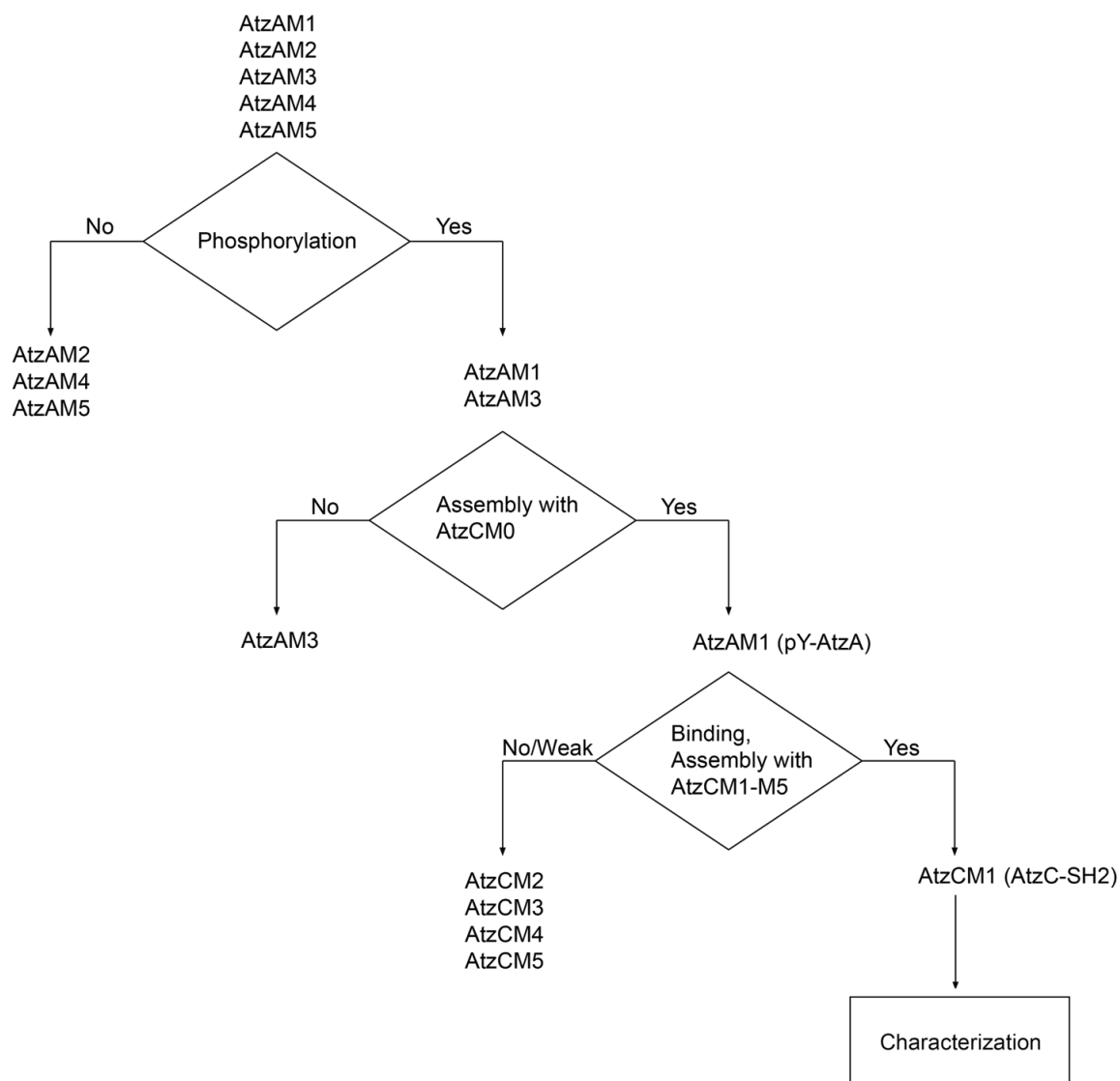
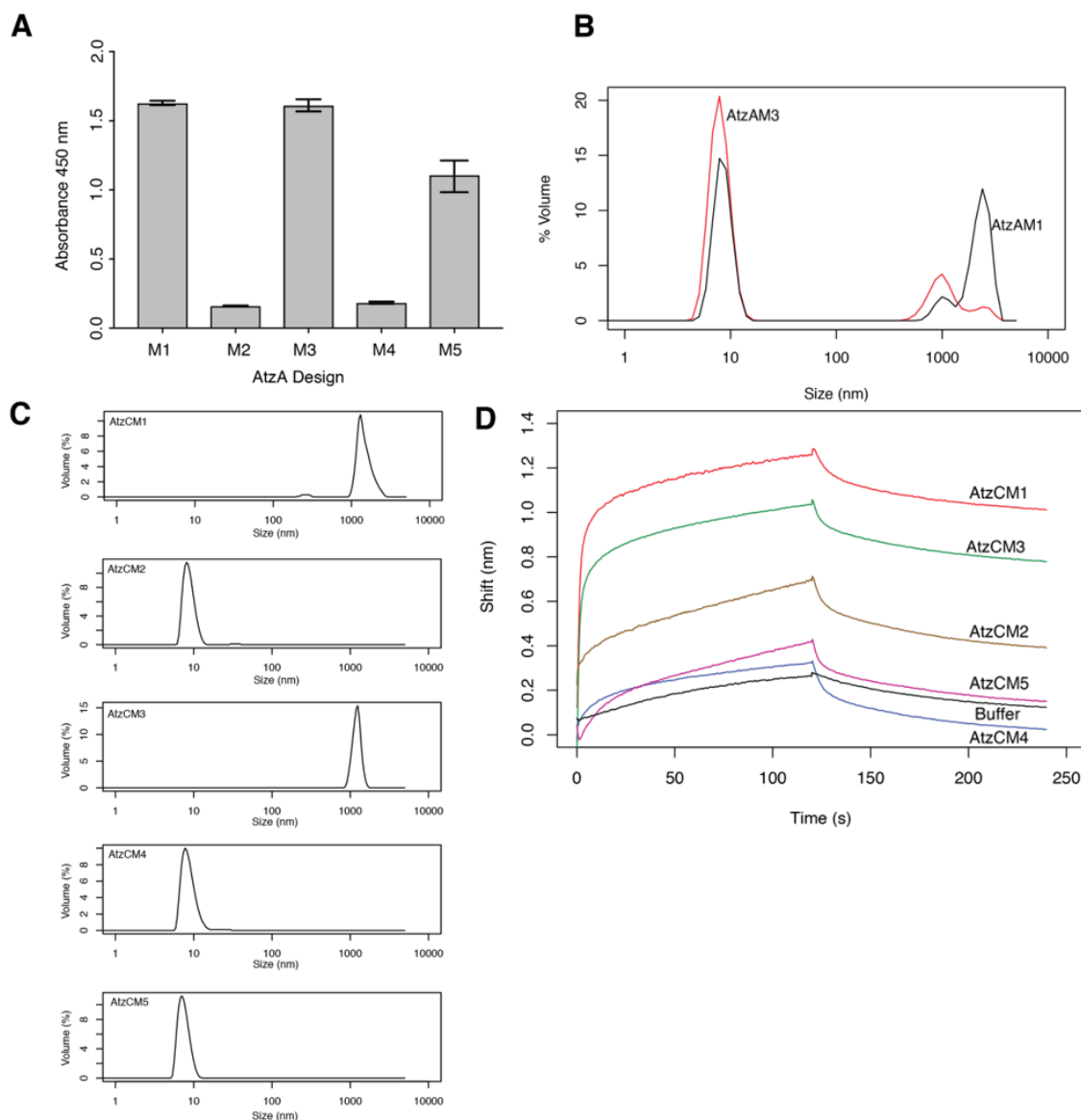


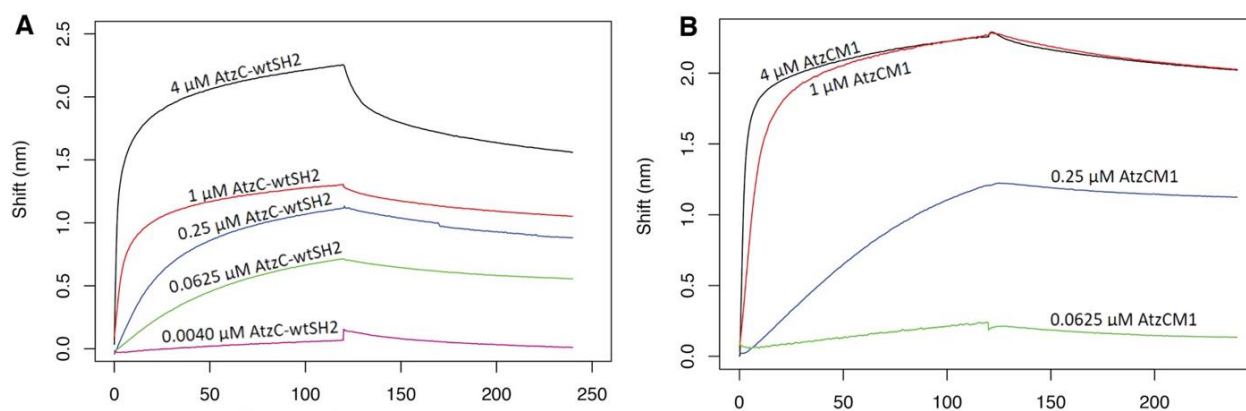
Figure 103-8. Supplementary Figure 8. Experimental selection process for pY-AtzA and AtzC-SH2.

Five N-terminal SH2 binding peptide AtzA fusions (AtzAM1-AtzAM5) and five C-terminal SH2 binding domain AtzC fusions (AtzCM1-AtzCM5) were selected, cloned, expressed, and purified. AtzAM1-M5 were screened for having the ability to be phosphorylated via ELISA with anti-phosphotyrosine. Only two AtzA designs, AtzAM1 and AtzAM3, showed strong phosphorylation. The ability for assembly formation to occur with a direct C-terminal SH2 binding domain AtzC fusion (no mutations; AtzCM0) was used to select the best AtzA design. AtzAM1 was chosen for superior assembly formation ability, becoming pY-AtzA. The five AtzC designs AtzCM1-AtzCM5 were screened for the ability to effectively bind and assemble with pY-AtzA. The combination of pY-AtzA and AtzCM1 (which we call AtzC-SH2) showed the strongest binding and the most robust assembly formation. This pair was then chosen for further characterization.



**Figure 104-8.** Supplementary Figure 9. Experimental selection of AtzA, AtzC subunits for characterization.

(A) ELISA screening of AtzA designs to determine phosphorylation levels. (B) DLS size distribution of AtzA designs with AtzCM0. (C) DLS size distribution of AtzC-SH2 designs with pY-AtzA. Samples prepared at 3  $\mu$ M pY-AtzA, 2  $\mu$ M AtzC-SH2 design. Only AtzCM1 and AtzCM3 showed assembly formation with pY-AtzA. Volume distribution reported. (D) BLI binding traces of AtzC-SH2 designs with pY-AtzA. AtzC-SH2 designs were screened for binding with BLI, using pY-AtzA as the load. Out of all AtzC-SH2 designs prepared, AtzCM1 had the highest binding affinity to pY-AtzA. Based on the assembly formation and binding data, AtzCM1 was chosen for further investigation.



**Figure 105-8.** Supplementary Figure 10. Bi-layer interferometry (BLI) binding profiles of AtzC wildtype SH2 fusion (AtzC-wtSH2) and AtzC superbinder SH2 fusion (AtzC-SH2) to phosphorylated SH2 binding peptide AtzA fusion (pY-AtzA).

(A) Binding profile of AtzC-wtSH2 to pY-AtzA. PY-AtzA was loaded onto the biosensor via a streptavidin-biotin interaction. AtzC-wtSH2 was flowed into the sample.  $K_D = 41.79 \pm 0.32$  nM. (B) Binding profile of AtzCM1 (superbinder) to pY-AtzA. PY-AtzA was loaded onto the biosensor via a streptavidin-biotin interaction. AtzC-SH2 was flowed into the sample.  $K_D = 7.67 \pm 0.52$  nM.

```

AtzCM0 MSKDFDLIIRNAYLSEKDSVYDIGIVGDRIIKIEAKIEGTVKDEIDAKGNLVSPGFVDAH 60
AtzCM1 MSKDFDLIIRNAYLSEKDSVYDIGIVGDRIIKIEAKIEGTVKDEIDAKGNLVSPGFVDAH 60
AtzCM2 MSKDFDLIIRNAYLSEKDSVYDIGIVGDRIIKIEAKIEGTVKDEIDAKGNLVSPGFVDAH 60
AtzCM3 MSKDFDLIIRNAYLSEKDSVYDIGIVGDRIIKIEAKIEGTVGDEIDAKGNLVSPGFVDAH 60
AtzCM4 MSKDFDLIIRNAYLSEKDSVYDIGIVGDRIIKIEAKIEGTVKDEIDAKGNLVSPGFVDAH 60
AtzCM5 MSKDFDLIIRNAYLSEKDSVYDIGIVGDRIIKIEAKIEGTVKDEIDAKGNLVSPGFVDAH 60

AtzCM0 THMDKSFTSTGERLPKFWSRPYTRDAAIEDGLKYYKNATHEEIKRHVIEHAHMQVLHGTL 120
AtzCM1 THMDKSFTSTGERLPKFWSRPYTRDAAIEDGLKYYKNATHEEIKRHVIEHAHMQVLHGTL 120
AtzCM2 THMDKSFTSTGERLPKFWSRPYTRDAAIEDGLKYYKNATHEEIKRHVIEHAHMQVLHGTL 120
AtzCM3 THMDKSFTSTGERLPKFWSRPYTRDAAIEDGLKYYKNATHEEIKRHVIEHAHMQVLHGTL 120
AtzCM4 THMDKSFTSTGERLPKFWSRPYTRDAAIEDGLKYYKNATHEEIKRHVIEHAHMQVLHGTL 120
AtzCM5 THMDKSFTSTGERLPKFWSRPYTRDAAIEDGLKYYKNATHEEIKRHVIEHAHMQVLHGTL 120

AtzCM0 YTRTHVDVDSVAKTKAVEAVLEAKEELKDLIDIQVVAFAQSGFFVDLESESLIRKSLDMG 180
AtzCM1 YTRTHVDVDSVAKTKAVEAVLEAKEELKDLIDIQVVAFAQSGFFVDLESESLIRKSLDMG 180
AtzCM2 YTRTHVDVDSVAKTKAVEAVLEAKEELKDLIDIQVVAFAQSGFFVDLESESLIRKSLDMG 180
AtzCM3 YTRTHVDVDSVAKTKAVEAVLEAKEELKDSIDIQVVAFAQSGFFVDLESESLIRKSLDMG 180
AtzCM4 YTRTHVDVDSVAKTKAVEAVLEAKEELKDLIDIQVVAFAQSGFFVDLESESLIRKSLDMG 180
AtzCM5 YTRTHVDVDSVAKTKAVEAVLEAKEELKDLIDIQVVAFAQSGFFVDLESESLIRKSLDMG 180

AtzCM0 CDLVGGVDPATRENNVEGSLDLCFKLAKEYDVIDIDYHIHDIGTVGVYSINRLAQKTIENG 240
AtzCM1 CDLVGGVDPATRENNVEGSLDLCFKLAKEYDVIDIDYHIHDIGTVGVYSINRLAQKTIENG 240
AtzCM2 CDLVGGVDPATRENNVEGSLDLCFKLAKEYDVIDIDYHIHDIGTVGVYSINRLAQKTIENG 240
AtzCM3 CDLVGGVDPATRENNVEGSLDLCFKLAKEYDVIDIDYHIHDIGTVGVYSINRLAQKTIENG 240
AtzCM4 CDLVGGVDPATRENNVEGSLDLCFKLAKEYDVIDIDYHIHDIGTVGVYSINRLAQKTIENG 240
AtzCM5 CDLVGGVDPATRENNVEGSLDLCFKLAKEYDVIDIDYHIHDIGTVGVYSINRLAQKTIENG 240

AtzCM0 YKGRVTTSHAWCFADAPSEWLDEAIPLYKDSGMKFVTCFSSPTPTMPVIKLEAGINLGC 300
AtzCM1 YKGRVTTSHAWCFADAPSEWLDEAIPLYKDSGMKFVTCFSSPTPTMPVIKLEAGINLGC 300
AtzCM2 YKGRVTTSHAWCFADAPSEWLDEAIPLYKDSGMKFVTCFSSPTPTMPVIKLEAGINLGC 300
AtzCM3 YKGRVTTSHAWCFADAPSEWLDEAIPLYKDSGMKFVTCFSSPTPTMPVIKLEAGINLGC 300
AtzCM4 YKGRVTTSHAWCFADAPSEWLDEAIPLYKDSGMKFVTCFSSPTPTMPVIKLEAGINLGC 300
AtzCM5 YKGRVTTSHAWCFADAPSEWLDEAIPLYKDSGMKFVTCFSSPTPTMPVIKLEAGINLGC 300

AtzCM0 ASDNIRDFWVPFGNGDMVQGALIIETQRLELKTNRDLGLIWKMITSEGARVLGIEKNYGIE 360
AtzCM1 ASDNIRDFWVPFGNGDMVQGALIIETQRLELKTNRDLGLIWKMITSEGARVLGIEKNYGIE 360
AtzCM2 ASDNIRDFWVPFGNGDMVQGALIIETQRLELKTNRDLGLIWKMITSEGARVLGIEKNYGIE 360
AtzCM3 ASDNIRDFWVPFGNGDMVQGALIIETQRLELKTNRDLGLIWKMITSEGARVLGIEKNYGIE 360
AtzCM4 ASDNIRDFWVPFGNGDMVQGALIIETQRLELKTNRDLGLIWKMITSEGARVLGIEKNYGIE 360
AtzCM5 ASDNIRDFWVPFGNGDMVQGALIIETQRLELKTNRDLGLIWKMITSEGARVLGIEKNYGIE 360

AtzCM0 VGKKADLVVLNSLSLSPQWAIIDQAKRLCVIKNGRIIVKDEVIVASIQAEWYFGKLGRKDA 420
AtzCM1 VGKKADLVVLNSLSLSPQWAIIDQAKRLCVIKNGRIIVKDEVIVAGGSAAEWYFGKLGRKDA 420
AtzCM2 VGKKADLVVLNSLSLSPQWAIIDQAKRLCVIKNGRIIVKDEVIVAGVAAEWYFGKLGRKDA 420
AtzCM3 VGKKADLVVLNSLSLSPQWAIIDQAKRLCVIKNGRIIVKDEVIIASGAAEWYFGKLGRKDA 420
AtzCM4 VGKKADLVVLNSLSLSPQWAIIDQAKRLCVIKNGAIIVKDEYIIAGGSAAEWYFGKLGRKDA 420
AtzCM5 VGKKADLVVLNSLSLSPQWAIIDQAKRLCVIKNGSICVKDEAIMASGSAAEWYFGKLGRKDA 420

```

Figure 106-8. Supplementary Figure 11a. Sequence alignment of AtzC-SH2 designs AtzCM0-AtzCM5.

AtzCM0	ERQLLSFGNPRGTFLIRESETVKGAYALSIRDWDDMKGDHVKHYLIRKLDNGGYYITTRA	480
AtzCM1	ERQLLSFGNPRGTFLIRESETVKGAYALSIRDWDDMKGDHVKHYLIRKLDNGGYYITTRA	480
AtzCM2	ERQLLSFGNPRGTFLIRESETVKGAYALSIRDWDDMKGDHVKHYLIRKLDNGGYYITTRA	480
AtzCM3	ERQLLSFGNPRGTFLIRESETVKGAYALSIRDWDDMKGDHVKHYLIRKLDNGGYYITTRA	480
AtzCM4	ERQLLSFGNPRGTFLIRESETVKGAYALSIRDWDDMKGDHVKHYLIRKLDNGGYYITTRA	480
AtzCM5	ERQLLSFGNPRGTFLIRESETVKGAYALSIRDWDDMKGDHVKHYLIRKLDNGGYYITTRA	480
AtzCM0	QFETLQQLVQHYSEARAAGLSSRLVVP SHKLE	517
AtzCM1	QFETLQQLVQHYSEARAAGLSSRLVVP SHKLE	517
AtzCM2	QFETLQQLVQHYSEARAAGLSSRLVVP SHKLE	517
AtzCM3	QFETLQQLVQHYSEARAAGLSSRLVVP SHKLE	517
AtzCM4	QFETLQQLVQHYSEARAAGLSSRLVVP SHKLE	517
AtzCM5	QFETLQQLVQHYSEARAAGLSSRLVVP SHKLE	517

**Figure 107-8.** Supplementary Figure 11b. Sequence alignment of AtzC-SH2 designs AtzCM0-AtzCM5 (con't).

Sequence alignment of AtzC-SH2 designs prepared. AtzCM0 is a direct fusion of AtzC and superbinder SH2 domain without mutations. Mutations made are highlighted in black or grey (similar residues). The red box highlights the region where the superbinder SH2 domain is located.

AtzAM0	MGSSHHHHHSSGLVPRGSHMEPQYEEIFNYQTLSIQHGTLVTMDQYRRVLGDSWVHVQD	60
AtzAM1	MGSSHHHHHSSGLVPRGSHMEPQYEEIFNYQTLSIQHGTLVTMDQYRRVLGDSWVHVQD	60
AtzAM2	MGSSHHHHHSSGLVPRGSHMEPQYEEIFNYGGLSIQHGTLVTMDQYRRVLGDSWVHVQD	60
AtzAM3	MGSSHHHHHSSGLVPRGSHMEPQYEEIFNYGGLSIQHGTLVTMDQYRRVLGDSWVHVQD	60
AtzAM4	MGSSHHHHHSSGLVPRGSHMEPQYEEIFDYGGLSIQHGTLVTMDQYRRVLGDSWVHVQD	60
AtzAM5	MGSSHHHHHSSGLVPRGSHMEPQYEEIFDYGTLSIQHGTLVTMDQYRRVLGDSWVHVQD	60
AtzAM0	GRIVALGVHAESVPPPADRVIDARGKVLPFGFINAHTHVNQILLRGGPSHGRQFYDWLFN	120
AtzAM1	GRIVALGVHAESVPPPADRVIDARGKVLPFGFINAHTHVNQILLRGGPSHGRQFYDWLFN	120
AtzAM2	GRIVALGVHAESVPPPADRVIDARGKVLPFGFINAHTHVNQILLRGGPSHGRQFYDWLFN	120
AtzAM3	GRIVALGVHAESVPPPADRVIDARGKVLPFGFINAHTHVNQILLRGGPSHGRQFYDWLFN	120
AtzAM4	GRIVALGVHAESVPPPADRVIDARGKVLPFGFINAHTHVNQILLRGGPSHGRQFYDWLFN	120
AtzAM5	GRIVALGVHAESVPPPADRVIDARGKVLPFGFINAHTHVNQILLRGGPSHGRQFYDWLFN	120
AtzAM0	VVYPGQKAMPEDVAVAVRLYCAEAVRSGITTINENADSAIYPGNI EAMAVYGEVGV RV	180
AtzAM1	VVYPGQKAMPEDVAVAVRLYCAEAVRSGITTINENADSAIYPGNI EAMAVYGEVGV RV	180
AtzAM2	VVYPGQKAMPEDVAVAVRLYCAEAVRSGITTINENADSAIYPGNI EAMAVYGEVGV RV	180
AtzAM3	VVYPGQKAMPEDVAVAVRLYCAEAVRSGITTINENADSAIYPGNI EAMAVYGEVGV RV	180
AtzAM4	VVYPGQKAMPEDVAVAVRLYCAEAVRSGITTINENADSAIYPGNI EAMAVYGEVGV RV	180
AtzAM5	VVYPGQKAMPEDVAVAVRLYCAEAVRSGITTINENADSAIYPGNI EAMAVYGEVGV RV	180
AtzAM0	VYARMFFDRMDGRIQGYVDALKARSPQVELCSIMEETAVAKDRITALSDQYHG TAGGRIS	240
AtzAM1	VYARMFFDRMDGRIQGYVDALKARSPQVELCSIMEETAVAKDRITALSDQYHG TAGGRIS	240
AtzAM2	VYARMFFDRMDGRIQGYVDALKARSPQVELCSIMEETAVAKDRITALSDQYHG TAGGRIS	240
AtzAM3	VYARMFFDRMDGRIQGYVDALKARSPQVELCSIMEETAVAKDRITALSDQYHG TAGGRIS	240
AtzAM4	VYARMFFDRMDGRIQGYVDALKARSPQVELCSIMEETAVAKDRITALSDQYHG TAGGRIS	240
AtzAM5	VYARMFFDRMDGRIQGYVDALKARSPQVELCSIMEETAVAKDRITALSDQYHG TAGGRIS	240
AtzAM0	VWPAPATTTAVTVEGMRWAQAFARDRAVMWTLHMAESDHDERIHGMSPAEYMECYGLLDE	300
AtzAM1	VWPAPATTTAVTVEGMRWAQAFARDRAVMWTLHMAESDHDERIHGMSPAEYMECYGLLDE	300
AtzAM2	VWPAPATTTAVTVEGMRWAQAFARDRAVMWTLHMAESDHDERIHGMSPAEYMECYGLLDE	300
AtzAM3	VWPAPATTTAVTVEGMRWAQAFARDRAVMWTLHMAESDHDERIHGMSPAEYMECYGLLDE	300
AtzAM4	VWPAPATTTAVTVEGMRWAQAFARDRAVMWTLHMAESDHDERIHGMSPAEYMECYGLLDE	300
AtzAM5	VWPAPATTTAVTVEGMRWAQAFARDRAVMWTLHMAESDHDERIHGMSPAEYMECYGLLDE	300
AtzAM0	RLQVAHCYVFDRKDVRLLRHNVKVASQVVSNAYLGSVAVPPEMVERGMAVGIGTDNGN	360
AtzAM1	RLQVAHCYVFDRKDVRLLRHNVKVASQVVSNAYLGSVAVPPEMVERGMAVGIGTDNGN	360
AtzAM2	RLQVAHCYVFDRKDVRLLRHNVKVASQVVSNAYLGSVAVPPEMVERGMAVGIGTDNGN	360
AtzAM3	RLQVAHCYVFDRKDVRLLRHNVKVASQVVSNAYLGSVAVPPEMVERGMAVGIGTDNGN	360
AtzAM4	RLQVAHCYVFDRKDVRLLRHNVKVASQVVSNAYLGSVAVPPEMVERGMAVGIGTDNGN	360
AtzAM5	RLQVAHCYVFDRKDVRLLRHNVKVASQVVSNAYLGSVAVPPEMVERGMAVGIGTDNGN	360
AtzAM0	SNDSVNMIGDMKFMAHIHRAVHRDADVLTPEKILEMATIDGARSLGMDHEIGSIETGKRA	420
AtzAM1	SNDSVNMIGDMKFMAHIHRAVHRDADVLTPEKILEMATIDGARSLGMDHEIGSIETGKRA	420
AtzAM2	SNDSVNMIGDMKFMAHIHRAVHRDADVLTPEKILEMATIDGARSLGMDHEIGSIETGKRA	420
AtzAM3	SNDSVNMIGDMKFMAHIHRAVHRDADVLTPEKILEMATIDGARSLGMDHEIGSIETGKRA	420
AtzAM4	SNDSVNMIGDMKFMAHIHRAVHRDADVLTPEKILEMATIDGARSLGMDHEIGSIETGKRA	420
AtzAM5	SNDSVNMIGDMKFMAHIHRAVHRDADVLTPEKILEMATIDGARSLGMDHEIGSIETGKRA	420
AtzAM0	DLILLDLRHPQTTPHHHLAATIVFQAYGNEVDTVLIDGNVVMENRRLSFLPPERELAFLE	480
AtzAM1	DLILLDLRHPQTTPHHHLAATIVFQAYGNEVDTVLIDGNVVMENRRLSFLPPERELAFLE	480
AtzAM2	DLILLDLRHPQTTPHHHLAATIVFQAYGNEVDTVLIDGNVVMENRRLSFLPPERELAFLE	480
AtzAM3	DLILLDLRHPQTTPHHHLAATIVFQAYGNEVDTVLIDGNVVMENRRLSFLPPERELAFLE	480
AtzAM4	DLILLDLRHPQTTPHHHLAATIVFQAYGNEVDTVLIDGNVVMENRRLSFLPPERELAFLE	480
AtzAM5	DLILLDLRHPQTTPHHHLAATIVFQAYGNEVDTVLIDGNVVMENRRLSFLPPERELAFLE	480
AtzAM0	EAQSRATAILQRANMVANPAWRSL	504
AtzAM1	EAQSRATAILQRANMVANPAWRSL	504
AtzAM2	EAQSRATAILQRANMVANPAWRSL	504
AtzAM3	EAQSRATAILQRANMVANPAWRSL	504
AtzAM4	EAQSRATAILQRANMVANPAWRSL	504
AtzAM5	EAQSRATAILQRANMVANPAWRSL	504

Figure 108-8. Supplementary Figure 12. Sequence alignment of pY-AtzA designs.

Sequence alignment of pY-AtzA designs prepared. AtzAM0 is a direct fusion of AtzA and SH2 binding peptide without mutations. Mutations made are shown in black. The red box indicates the SH2 recognition peptide sequence.

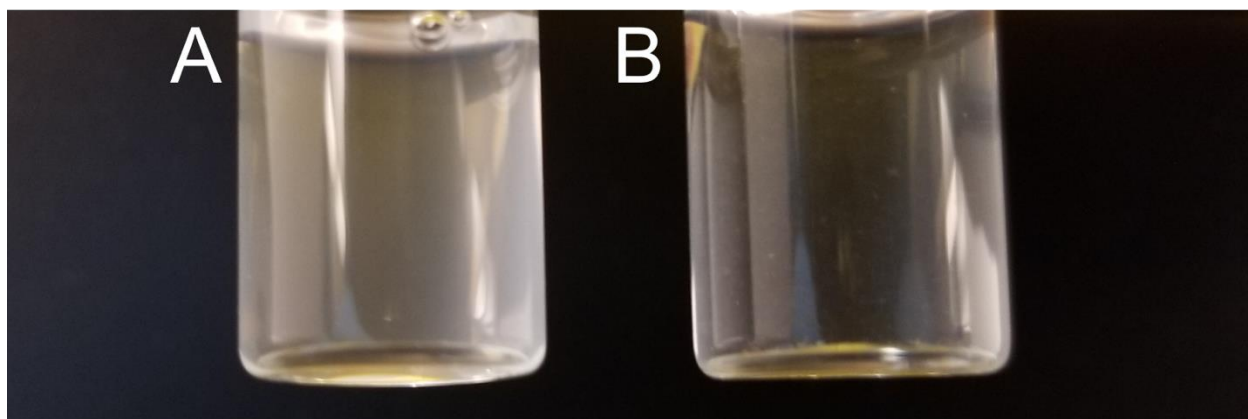
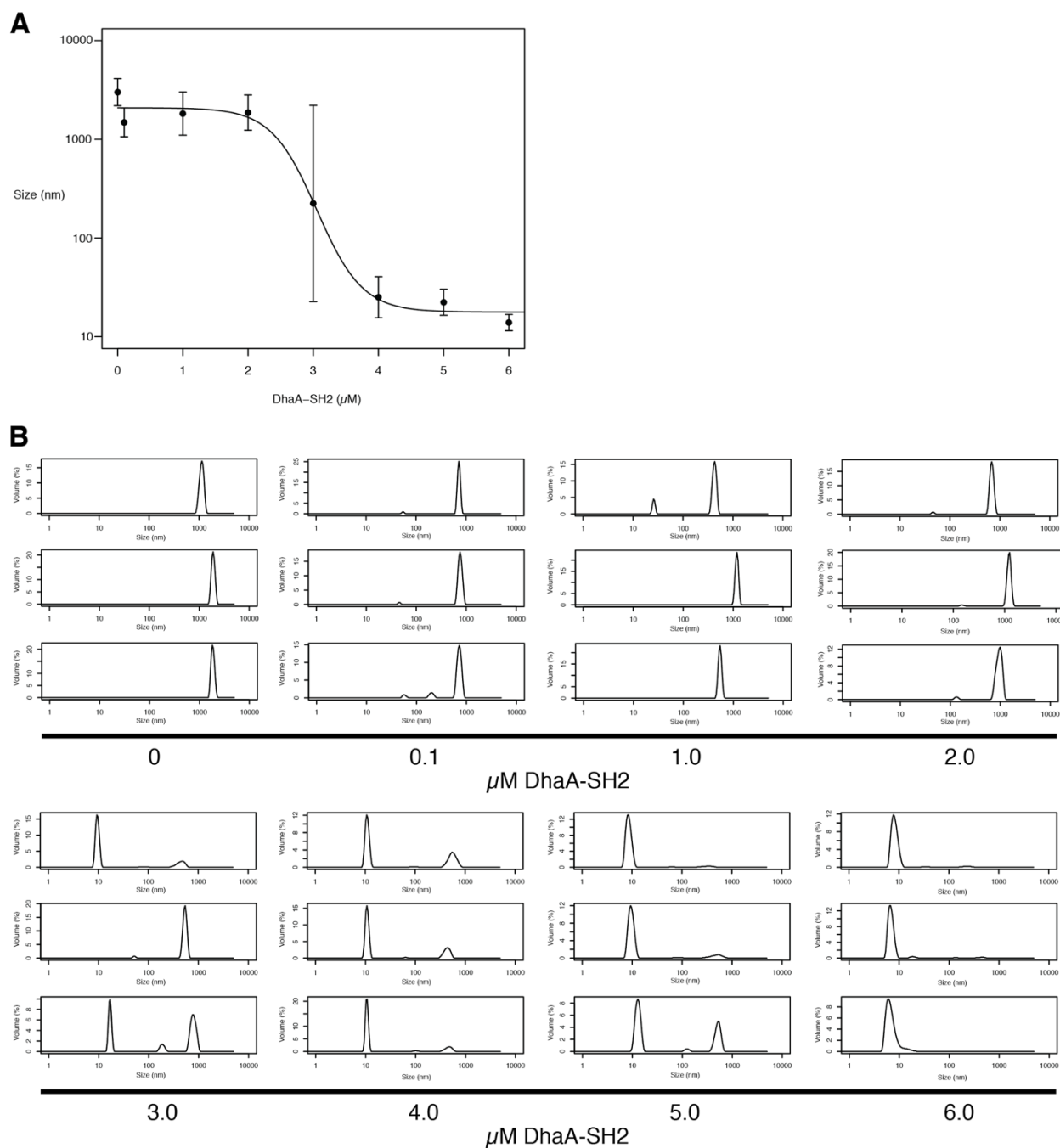


Figure 109-8, Supplementary Figure 13. Visual assembly turbidity.

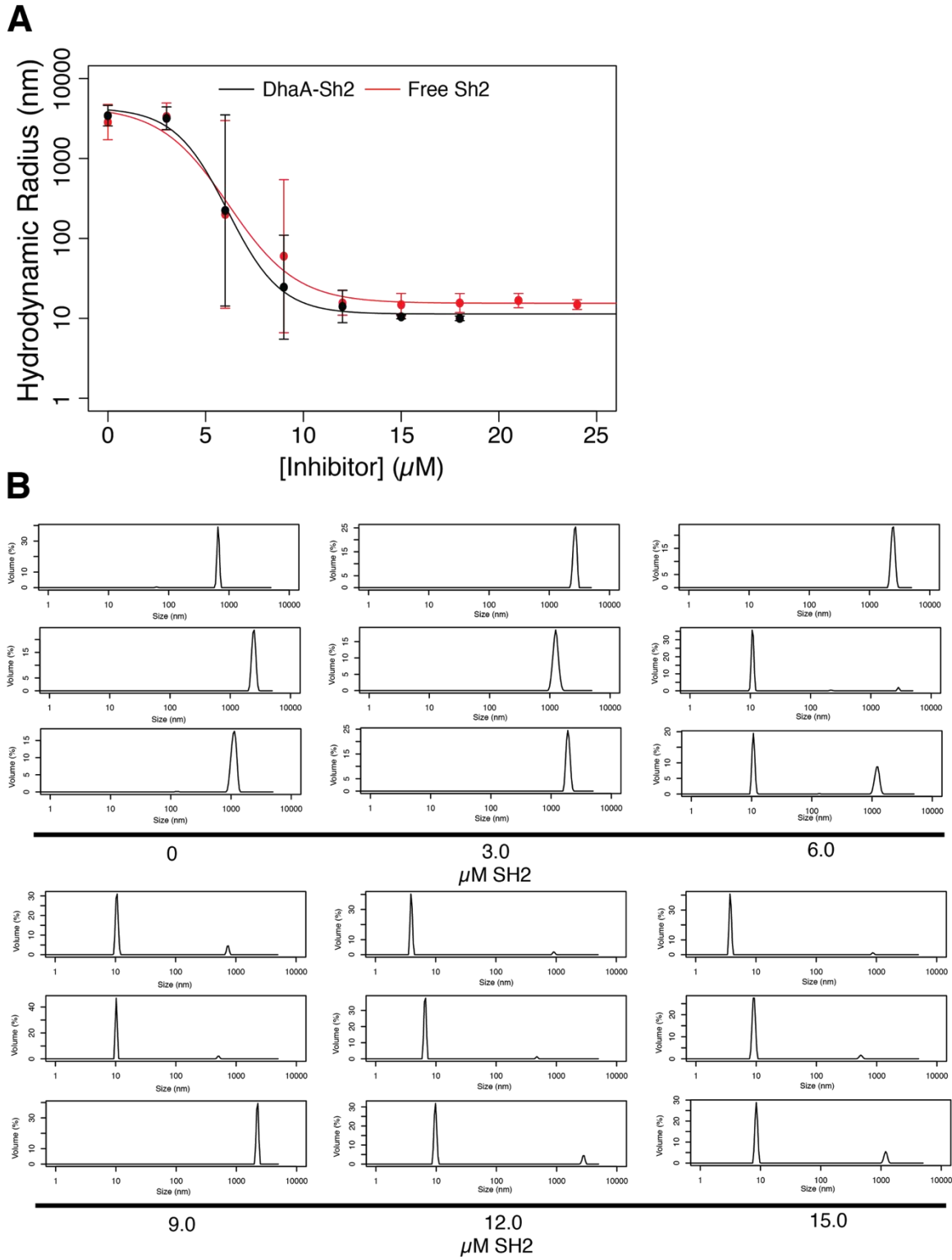
(A) 3  $\mu\text{M}$  pY-AtzAM1 and 2  $\mu\text{M}$  AtzCM1, shows a turbid solution that represents the assembly formed. (B) 3  $\mu\text{M}$  non-pY-AtzAM1 and 2  $\mu\text{M}$  AtzCM1, shows a clear solution with no assembly formation.



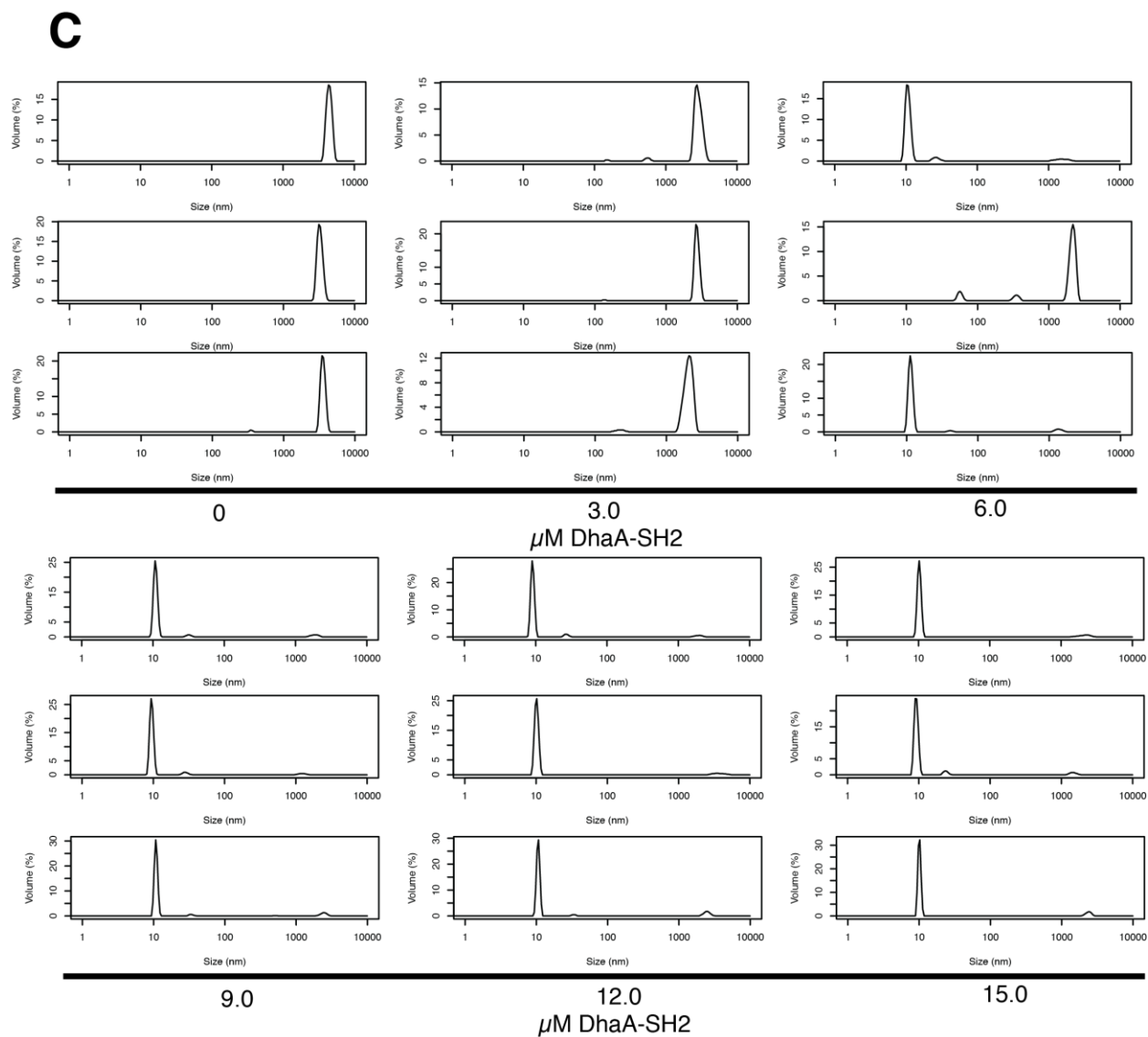
**Figure 110-8.** Supplementary Figure 14. Inhibition of assembly at 0.66  $\mu\text{M}$  AtzC-SH2, 1  $\mu\text{M}$  pY-AtzA, 0-6  $\mu\text{M}$  SH2-DhaA.

(A) Inhibition graph of SH2-DhaA on 0.66  $\mu\text{M}$  AtzC-SH2, 1  $\mu\text{M}$  pY-AtzA assembly. Size recorded represents most predominant DLS sizing peak. Data are presented as mean  $\pm$  1 standard deviation.  $\text{IC}_{50} = 3.05 \mu\text{M}$ . Adjusted  $R^2 = 0.98$ .

(B) DLS traces of assembly from 0 - 6  $\mu\text{M}$  SH2-DhaA. DLS traces are of triplicates.

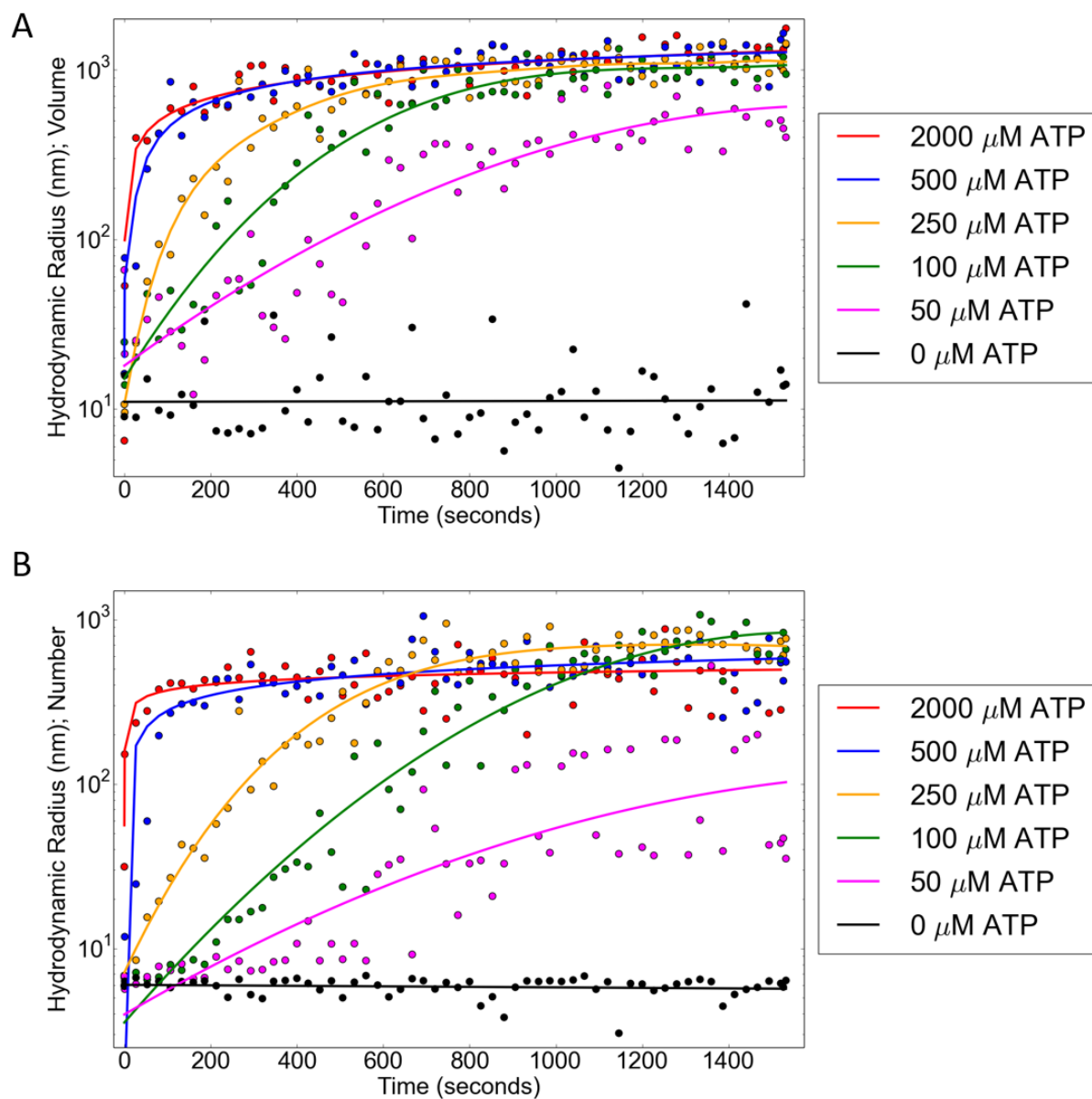


**Figure 111-8.** Supplementary Figure 15. Inhibition of assembly at 2  $\mu\text{M}$  AtzC-SH2, 3  $\mu\text{M}$  pY-AtzA with 0-15  $\mu\text{M}$  inhibitor.



**Figure 112-8.** Supplementary Figure 16. Inhibition of assembly at 2  $\mu\text{M}$  AtzC-SH2, 3  $\mu\text{M}$  pY-AtzA with 0-15  $\mu\text{M}$  inhibitor (con't).

All DLS traces were performed in triplicate (A) Inhibition graph of SH2-DhaA of 2  $\mu\text{M}$  AtzC-SH2, 3  $\mu\text{M}$  pY-AtzA assembly. Size recorded represents most predominant DLS sizing peak. Data are presented as mean  $\pm$  1 standard deviation.  $\text{IC}_{50}$  (SH2) = 6.18  $\mu\text{M}$ ,  $\text{IC}_{50}$  (SH2-DhaA) = 6.13  $\mu\text{M}$ . Adjusted  $R^2$  (SH2) = 0.97. Adjusted  $R^2$  (SH2-DhaA) = 0.99. (B) DLS traces of assembly from 0-15  $\mu\text{M}$  SH2. (C) DLS traces of assembly from 0-15  $\mu\text{M}$  SH2-DhaA.

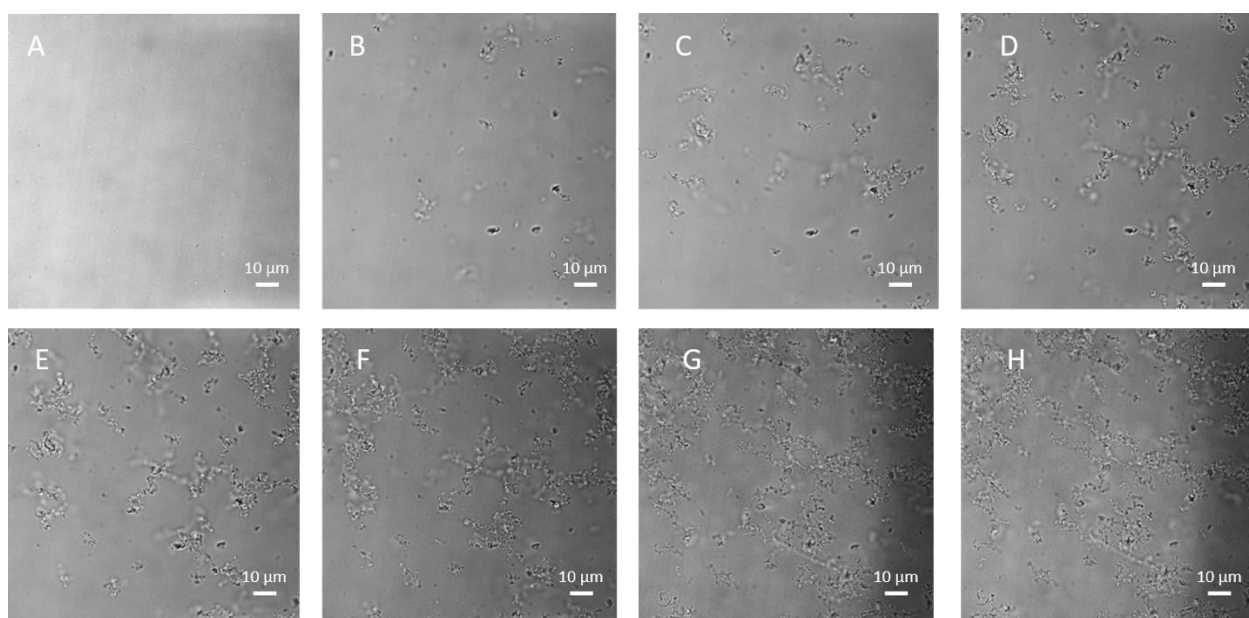


**Figure 113-8.** Supplementary Figure 17. Rate of assembly formation is dependent on ATP concentration.

(A) Volume mean of sample from 0 – 1500 sec. Each point represents average of triplicates. (B) Number mean of sample from 0 – 1500 sec. Each point represents average of triplicates. Curve fitting performed using sloping spline with smoothness parameter ( $p$ ) and adjusted  $R^2$  value given in Table S1 (highest concentration of ATP to lowest, starting from top to bottom at time 0 for both graphs).

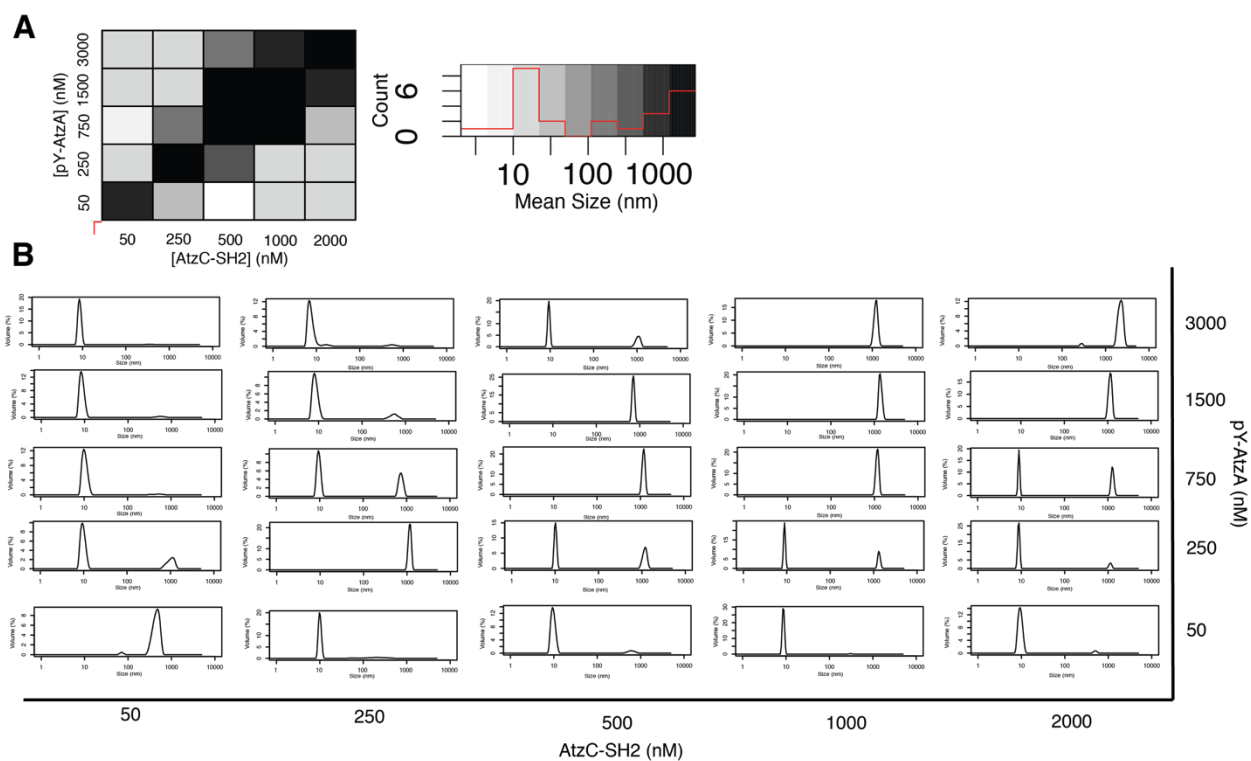
Table 16-8. Supplementary Table 2. Curve fitting data for Supplementary Figure 16.  
Adjusted R<sup>2</sup> and smoothing parameter ( $p$ ) value given for curve fitting done on assembly kinetics data.

Distribution	ATP $\mu$ M	Adjusted R-square	$p$
Vol	2000	0.7889	1.31E-05
Vol	500	0.888	1.31E-05
Vol	250	0.898	3.25E-08
Vol	100	0.9374	3.25E-08
Vol	50	0.867	3.25E-08
Vol	0	0.2638	0.000182922
Num	2000	0.5044	2.16E-05
Num	500	0.9303	2.16E-05
Num	250	0.9678	2.16E-05
Num	100	0.9543	2.16E-05
Num	50	0.7189	7.25E-09
Num	0	0.3338	0.000110956



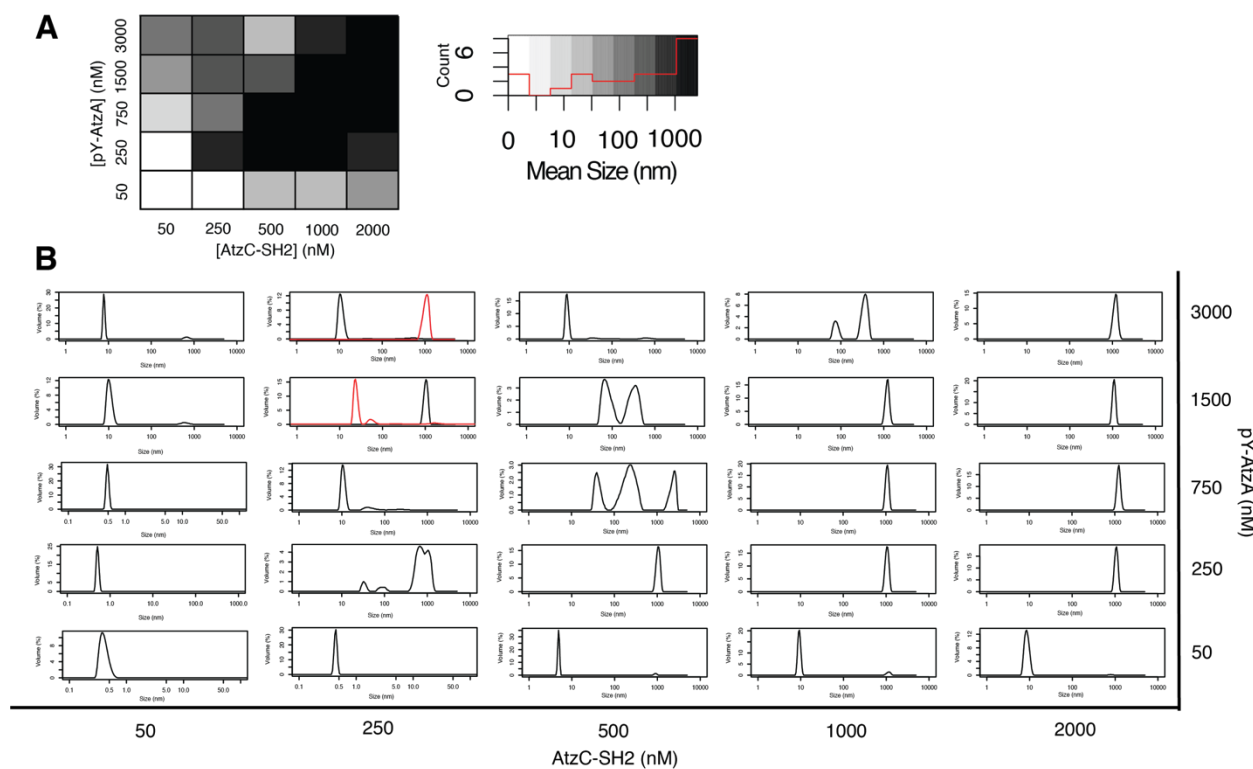
**Figure 114-8, Supplementary Figure 18.** Bright-field view of the assembly growing after the addition of Src kinase.

(A) 3 minutes after addition of Src kinase, no assemblies shown. (B) 14 minutes after addition of Src kinase, small assemblies shown. (C) 18 minutes after addition of Src kinase, small 10  $\mu\text{m}$  assemblies start to grow (D) 24 minutes after the addition of Src kinase, growth continues. (E) 30 minutes after addition of Src kinase, over 50  $\mu\text{m}$  size assemblies form. (F) 35 minutes after addition of Src kinase, 100  $\mu\text{m}$  size assemblies appear. (G) 40 minutes after addition of Src kinase, assemblies continue to grow. (H) 50 minutes after addition of Src kinase, assemblies have fully matured into fractal-like structures.



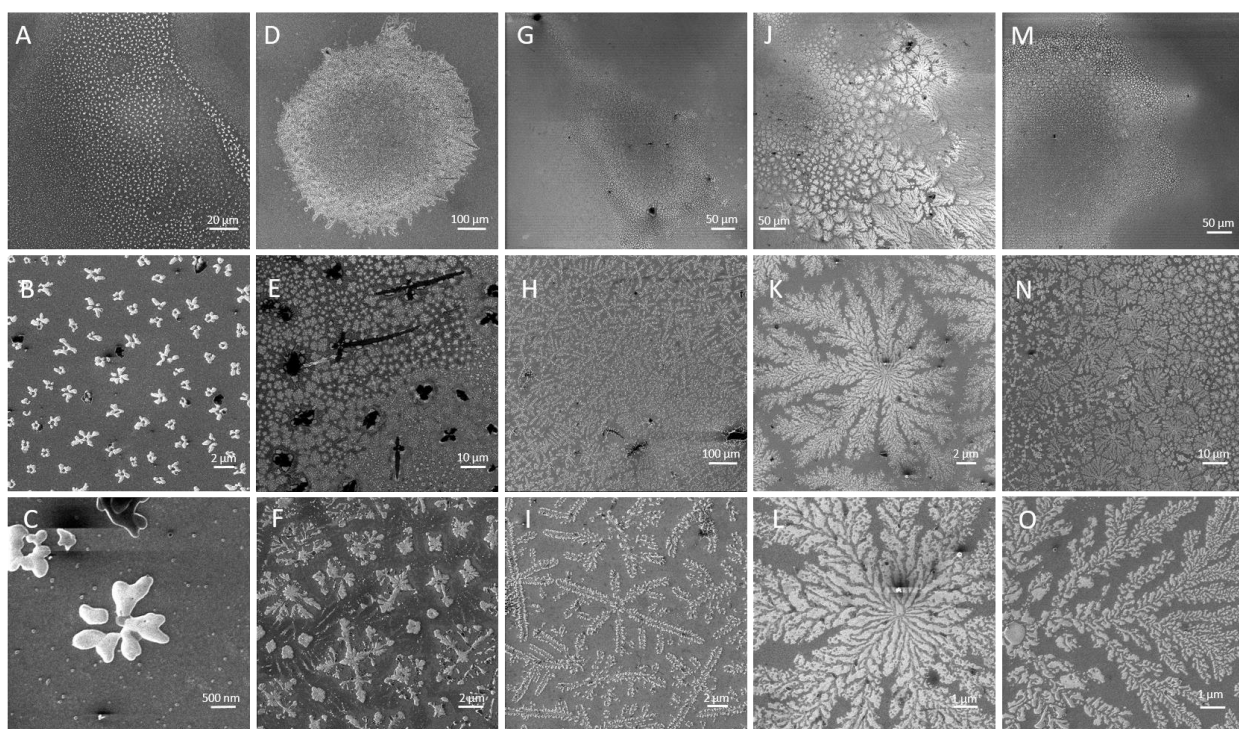
**Figure 115-8.** Supplementary Figure 19. Average size of particle formed by pY-AtzA and wild type AtzC-SH2.

(A) Heat map showing volume-weighted mean size of particles found from 50-3000 nM pY-AtzA and 50-2000 nM AtzC-SH2. Value shown is average of two physical samples. Histogram illustrates distribution of sizes found on heatmap. (B) Volume distributions of heat map. Distributions shown are representative of other traces in the sample.



**Figure 116-8.** Supplementary Figure 20. Average size of particle formed by pY-AtzA and super-binder AtzC-SH2.

(A) Heat map showing volume-weighted mean size of particles found from 50-3000 nM pY-AtzA and 50-2000 nM AtzC-SH2. Value shown is average of two physical samples. Histogram illustrates distribution of sizes found on heatmap. (B) Volume distributions of heat map. Distributions shown are representative of other traces in the sample.



**Figure 117-8.** Supplementary Figure 21. Helium ion microscopy (HIM) depict fractal-like assembly with increasing AtzA concentrations.

(A to C) 0.250  $\mu\text{M}$  AtzAM1 and 2  $\mu\text{M}$  AtzCM1. (D to F) 0.950  $\mu\text{M}$  AtzAM1 and 2  $\mu\text{M}$  AtzCM1 (G-I) 1.5  $\mu\text{M}$  AtzAM1 and 2  $\mu\text{M}$  AtzCM1. (J to L) 3  $\mu\text{M}$  AtzAM1 and 2  $\mu\text{M}$  AtzCM1. (M to O) 3  $\mu\text{M}$  AtzAM1 and 1  $\mu\text{M}$  AtzCM1.

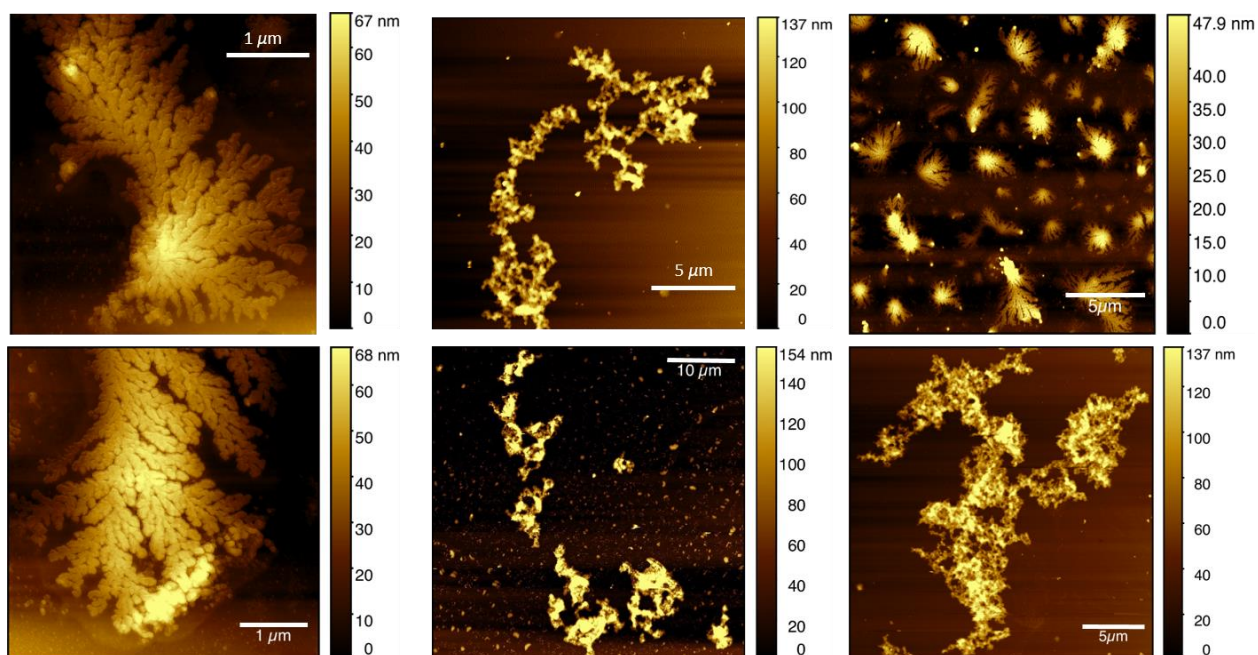
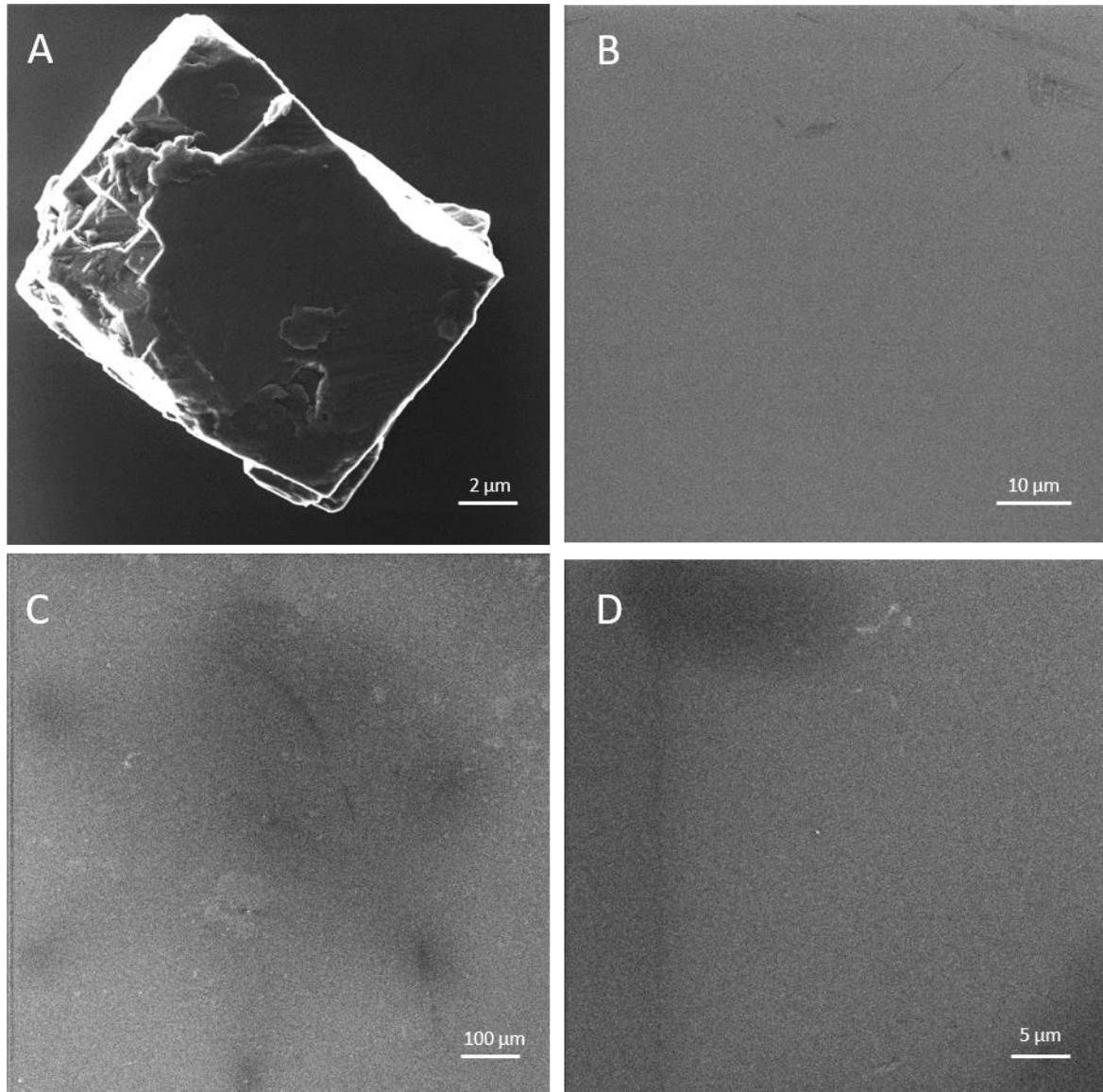


Figure 118-8. Supplementary Figure 22. Atomic Force Microscopy (AFM) images show fractal-like structures, fern-like, and petal-like structures, similar to HIM.



**Figure 119-8.** Supplementary Figure 23. Helium ion microscopy (HIM) buffer and non-phosphorylated controls preclude salt precipitation.

In order to determine that our proteins were forming fractal-like patterns and it was not salt inducing the patterns, a buffer and non-phosphorylated proteins sample controls were used to preclude salt precipitation. (A) Usual HIM square salt crystals on a glass surface. (B) Deposited HNG buffer (50 mM Hepes, 100 mM NaCl, 5% glycerol, pH 7.4, buffer proteins are stored in) on silicon wafer shows no structures on the surface. (C) 3  $\mu$ M non-pY-AtzAM1 and 2  $\mu$ M AtzCM1 control shows no fractal-like structures. (D) 3  $\mu$ M non-pY-AtzAM1 and 1  $\mu$ M AtzCM1 show no fractal-like structures. All controls demonstrate that fractal structures are formed by phosphorylated protein components.

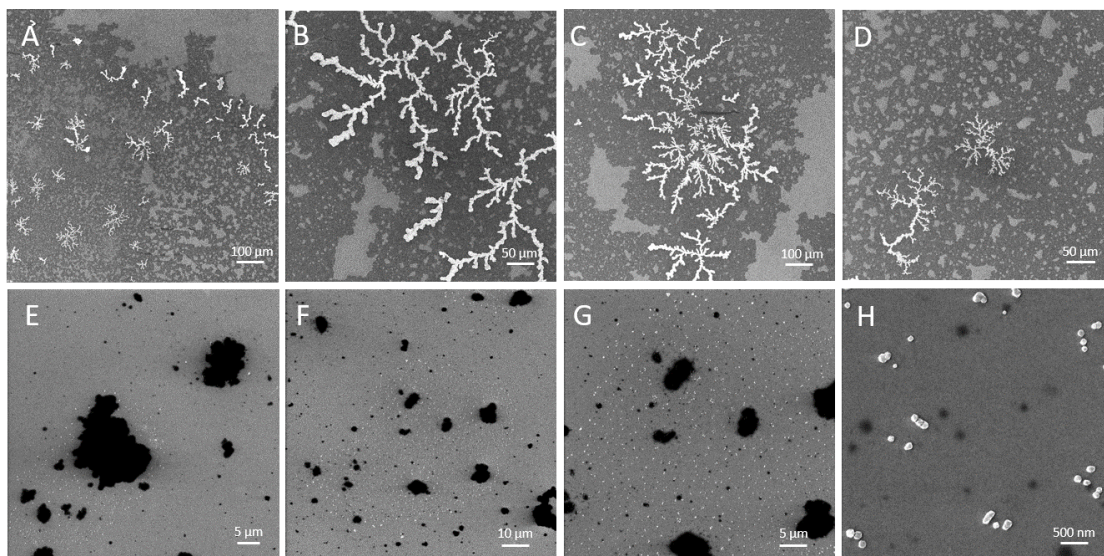
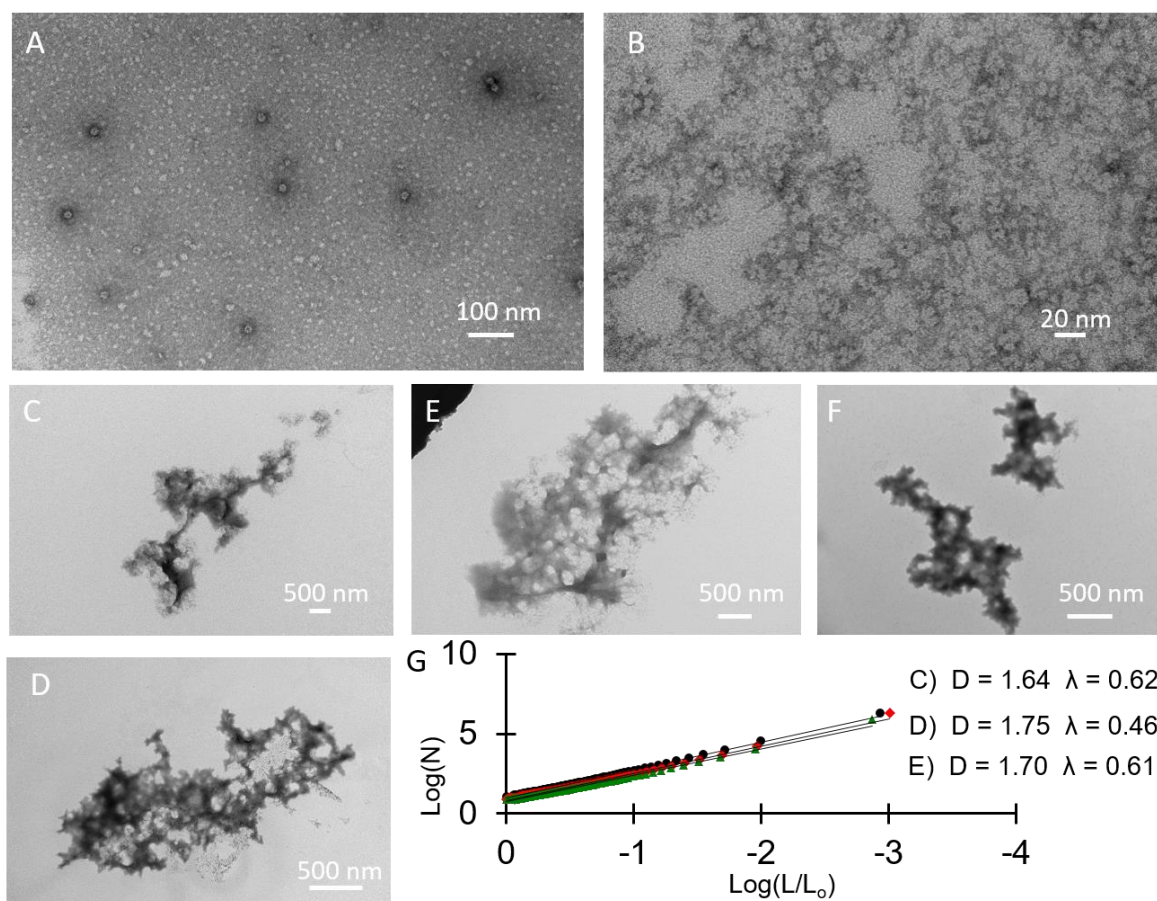


Figure 120-8, Supplementary Figure 24. Helium ion microscopy comparison of fractal assembly and globular assembly. HIM Images depict fractal-like assembly with 3 uM AtzAM1 and 2 uM AtzCM1 final concentrations (A to D), while the 3 uM AtzAM1-ExtendedLinker and 2 uM AtzCM1-ExtendedLinker final concentrations show both large and small globular shape proteins on the silicon surface (E to H).



**Figure 121-8.** Supplementary Figure 25. Transmission Electron Microscopy (TEM) depicts fractal-like assemblies in the phosphorylated samples while the non-phosphorylated samples depict individual proteins.

(A and B) ten-fold dilution of 3  $\mu\text{M}$  non-pY-AtzAM1 and 2  $\mu\text{M}$  AtzCM1, which shows the individual proteins. (C to F) Various assembly images of the ten-fold dilution of 3  $\mu\text{M}$  pY-AtzAM1 and 2  $\mu\text{M}$  AtzCM1 sample which form the fractal-like assembly consistently. (G) Image analysis (2D) using box counting yields the expected fractal dimension of  $\sim 1.7$  for the C, D, and E, TEM images.

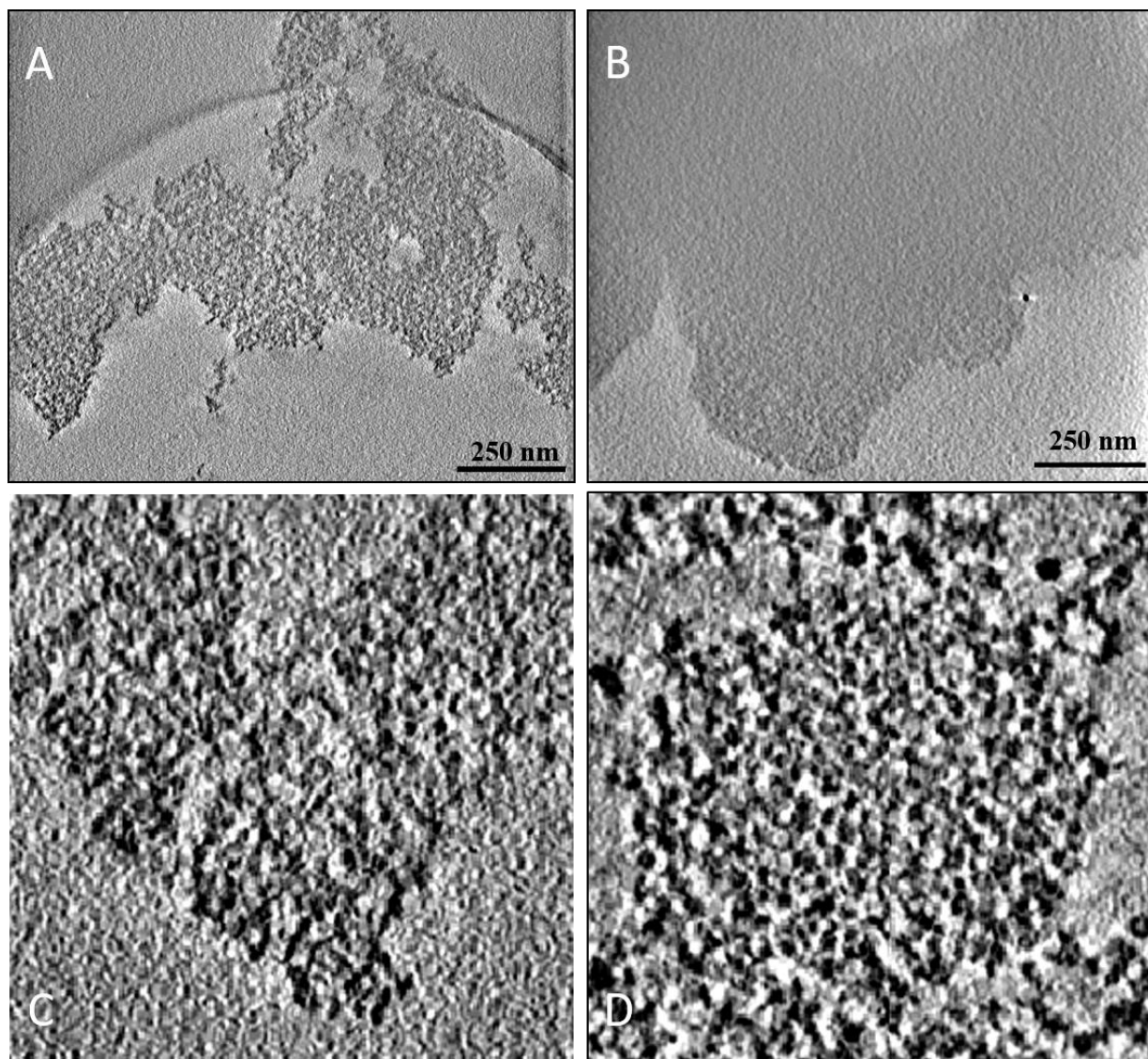
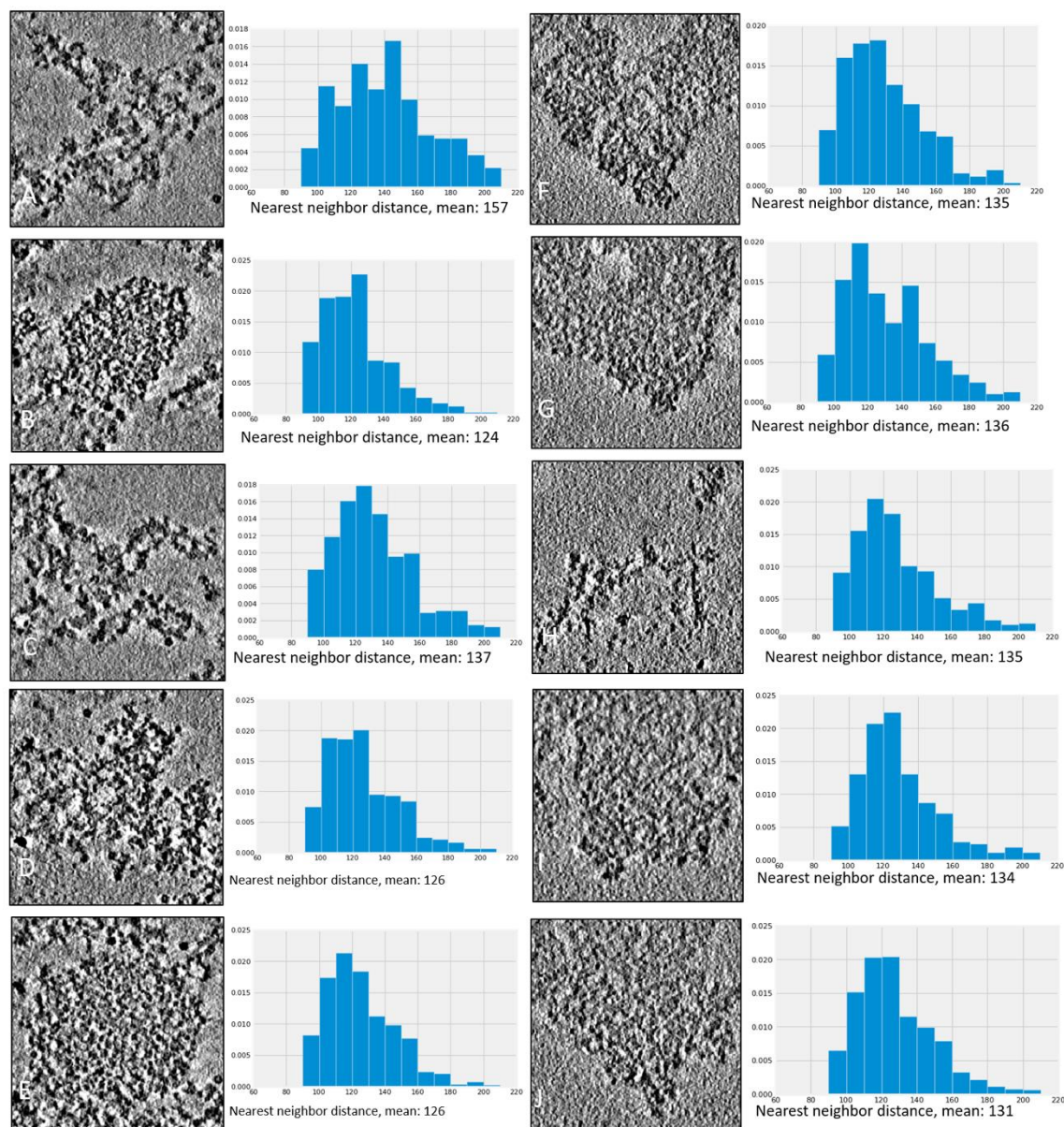


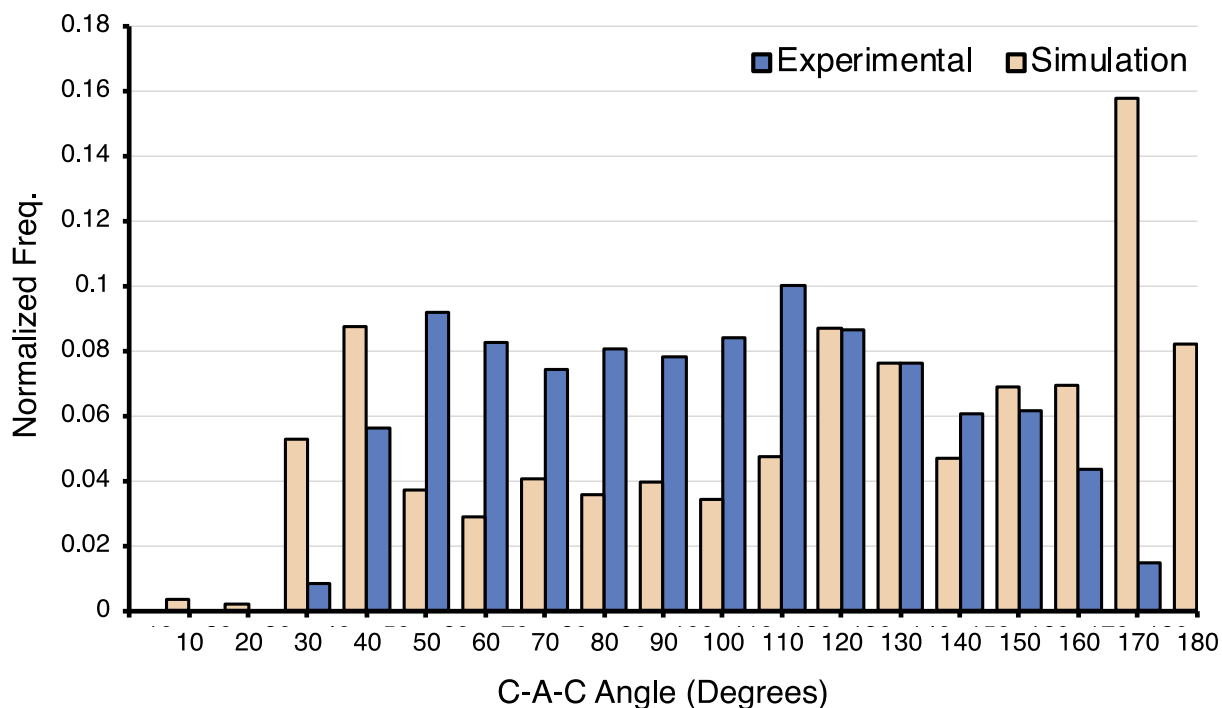
Figure 122-8. Supplementary Figure 26. Comparison of the fractal assembly CryoEM tomograms and the extended linker globular assemblies.

CryoEM tomograms of the fractal-like assemblies (A) and the extended linker assemblies (B) show a difference in the overall topology of the two different assemblies. Zoomed in versions of the images show representatives of a fractal assembly (C) and of a very dense and globular structure (D).



**Figure 123-8.** Supplementary Figure 27. Analysis of the fractal assembly CryoEM tomograms and the extended linker globular assemblies.

CryoEM tomograms of the GS-linker assemblies (A-E) and fractal assemblies (F-J) next to the calculated nearest neighbor distance and mean average distance are shown.



**Figure 124-8.** Supplementary Figure 28. Spatial angle comparison of Cryo-EM structure to simulation.

For the large assembly discussed in Figure 4, the angle formed from three proximal components (C-A-C) was calculated using experimentally derived Cryo-EM density fits and the geometric centers of the simulated assembly structures. While both the experimental and simulated assemblies distributions are relatively flat, simulated assemblies show a marked preference for values 30, 170, and 180 compared to the experimental assemblies. The disagreement between the distributions at 170 and 180 may arise due to lack of flexibility in the simulated assemblies which effectively decreases sampling of non-linear connections due to detected steric clashes in a growing assembly. The experimentally derived angular distribution is significantly more sensitive to the assignment of density to individual components compared to inter-component distances. The uncertainty in placement at this resolution ( $\sim 40\text{\AA}$ ) may also contribute to the observed differences between experimentally derived and calculated angular distributions more than the errors in distance distributions in Fig.4G)

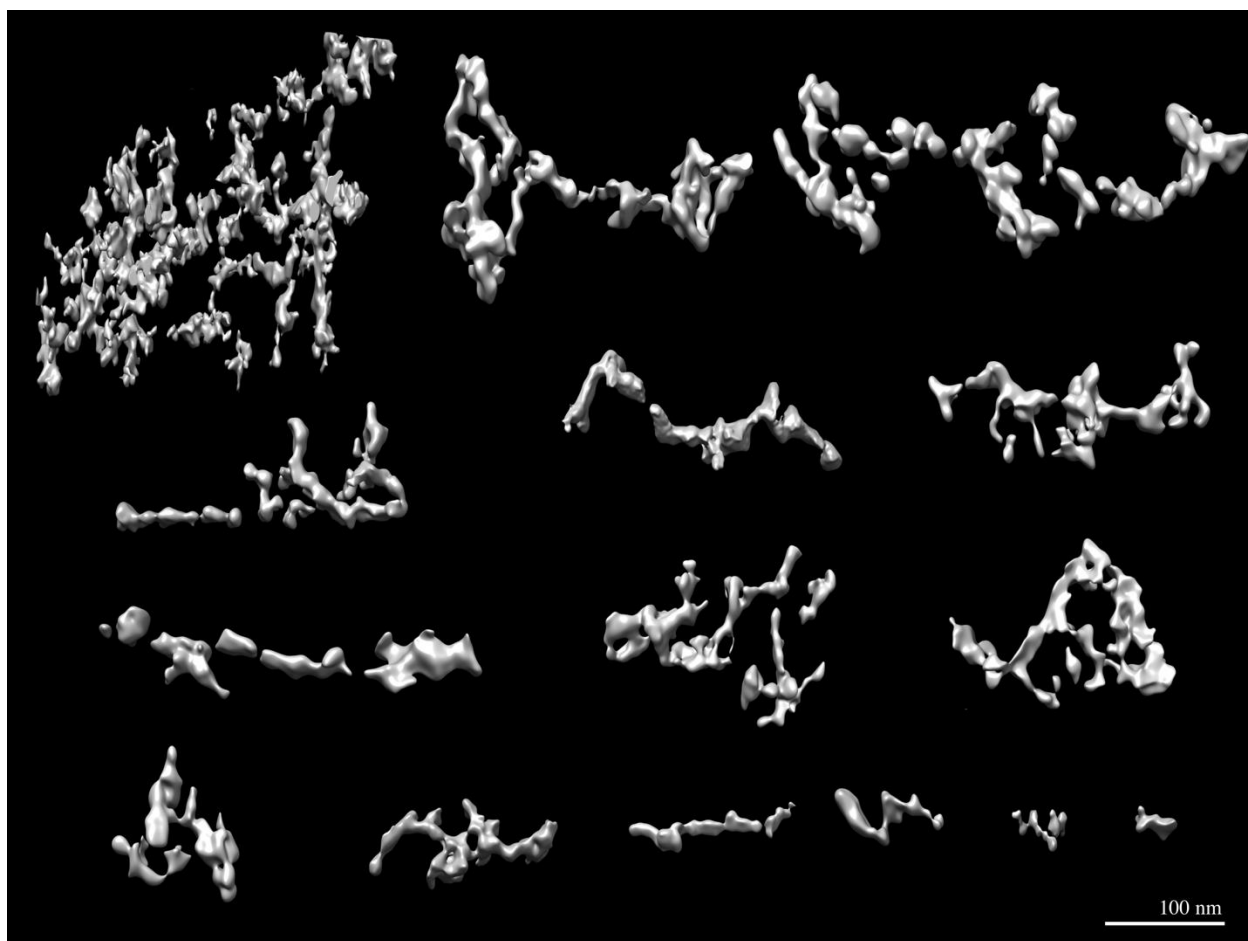
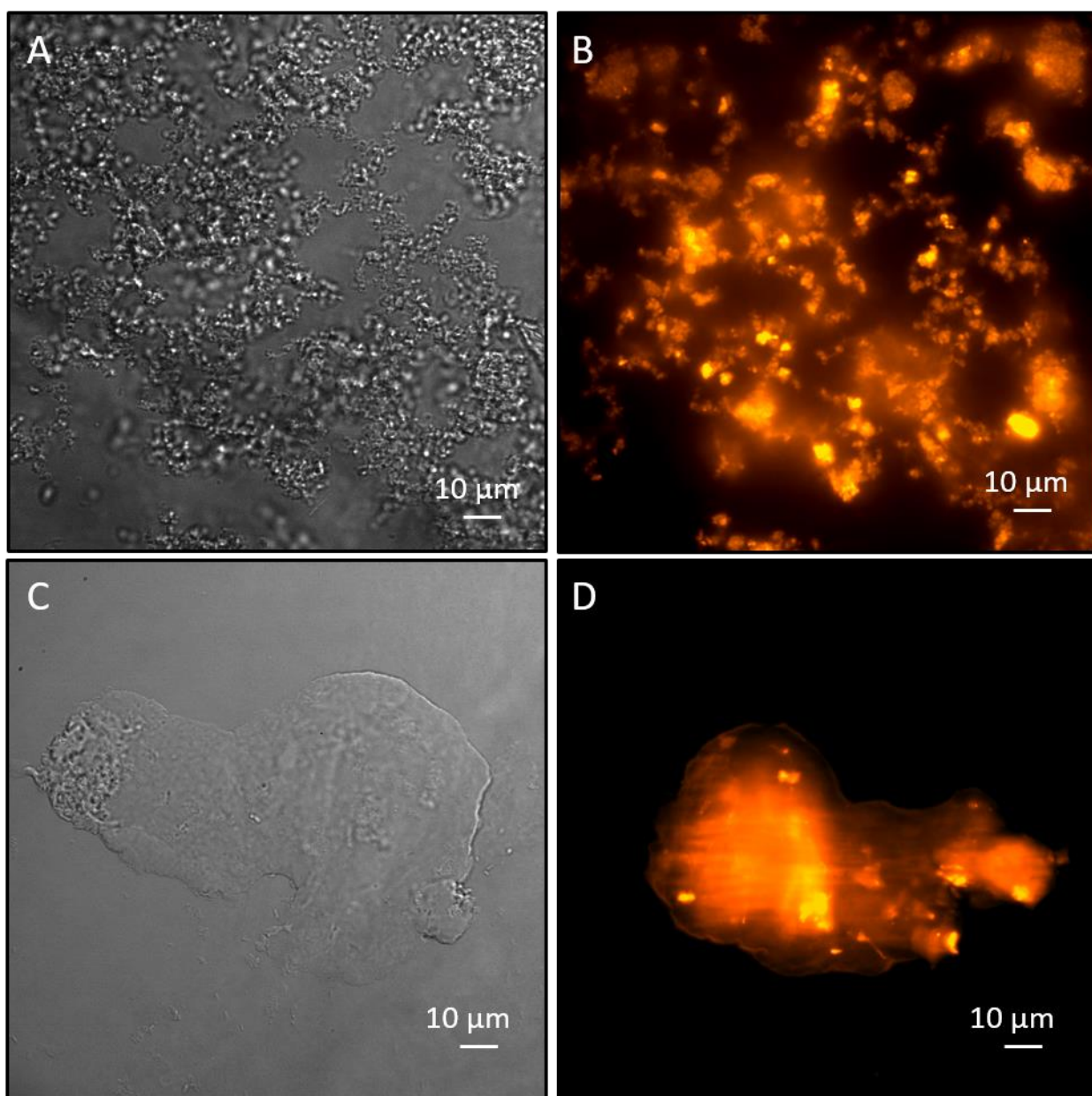
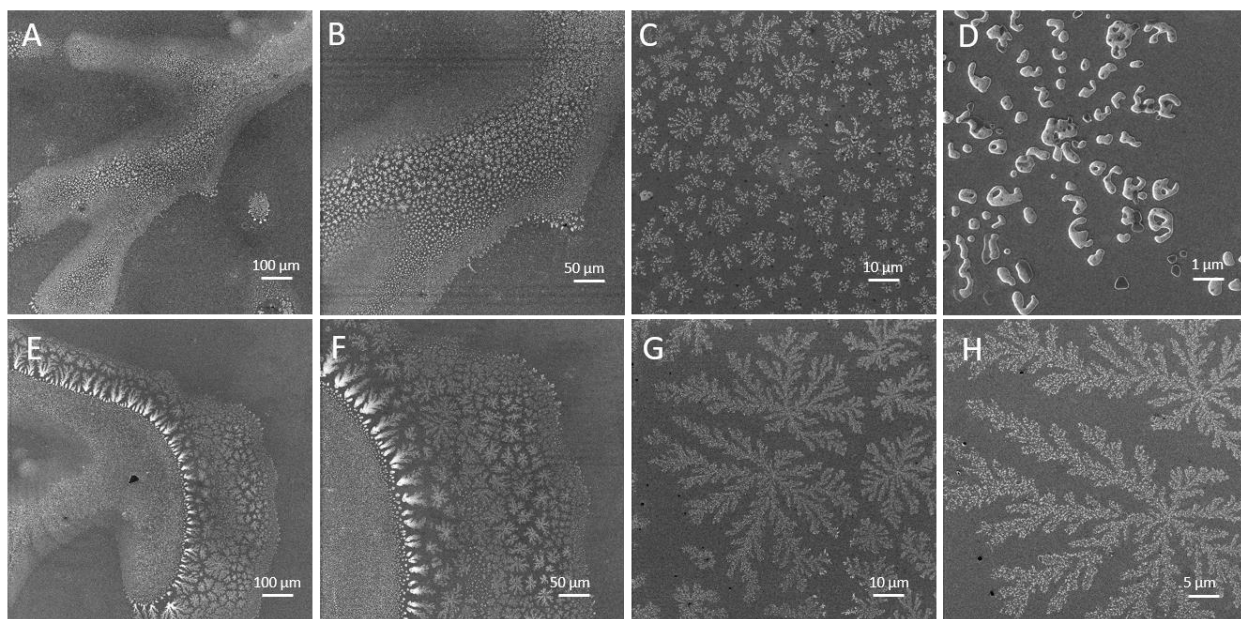


Figure 125-8. Supplementary Figure 29. Isosurface views of the assembly tomograms, from large to small.



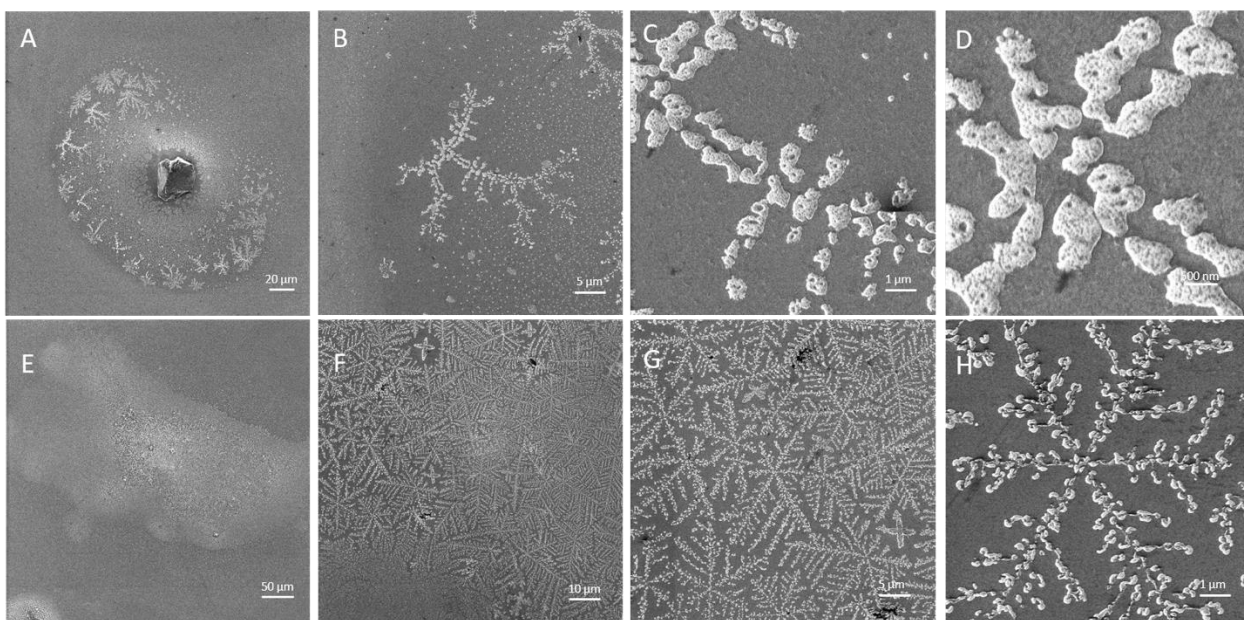
**Figure 126-8.** Supplementary Figure 30. Fluorescence microscopy and bright-field images of the 4-component assembly (AtzAM1, AtzCM1, ProteinA-SH2, and antibody, along with extended linker versions of AtzA and AtzC).

Images confirm incorporation of IgG-Antibody-Alexa Fluor 568 into assemblies. (A) Fractal assembly in DIC and (B) fluorescent image of fractal indicating incorporation of antibody into assembly. (C) Globular assembly in DIC and (D) fluorescent images of globular assembly indicating incorporation of antibody into assembly. The depiction of a fractal and globular topology is easily distinguishable in these images.



**Figure 127-8.** Supplementary Figure 31. Helium ion microscopy (HIM) images depict fractal-like assembly with 3  $\mu\text{M}$  AtzAM1, 1  $\mu\text{M}$  AtzBSH2, 1  $\mu\text{M}$  AtzCM1 final protein concentrations.

(A to D) Various views of the fractal-like 3-component assembly are shown.

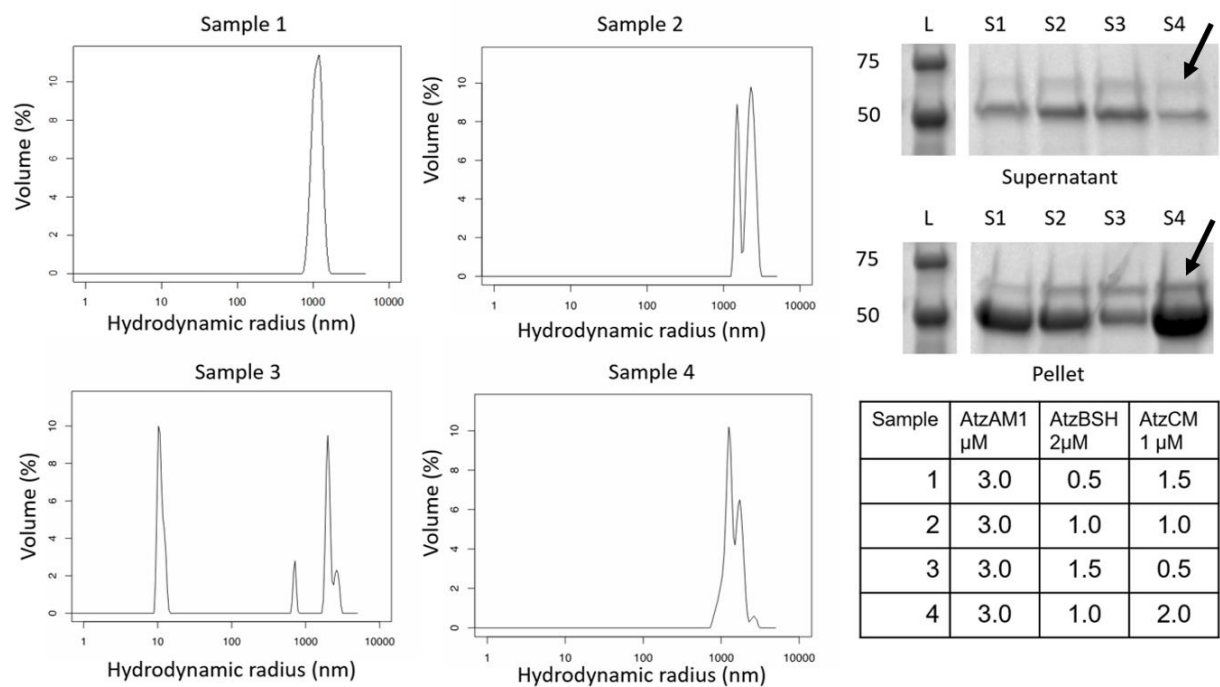


**Figure 128-8.** Supplementary Figure 32. Helium ion microscopy (HIM) images depict fractal-like assembly with 3  $\mu$ M AtzAM1, 1  $\mu$ M AtzBSH2, 2  $\mu$ M AtzCM1 final concentrations.

(A to H) Various views of the 3-component assembly with fractal-like structures are shown.

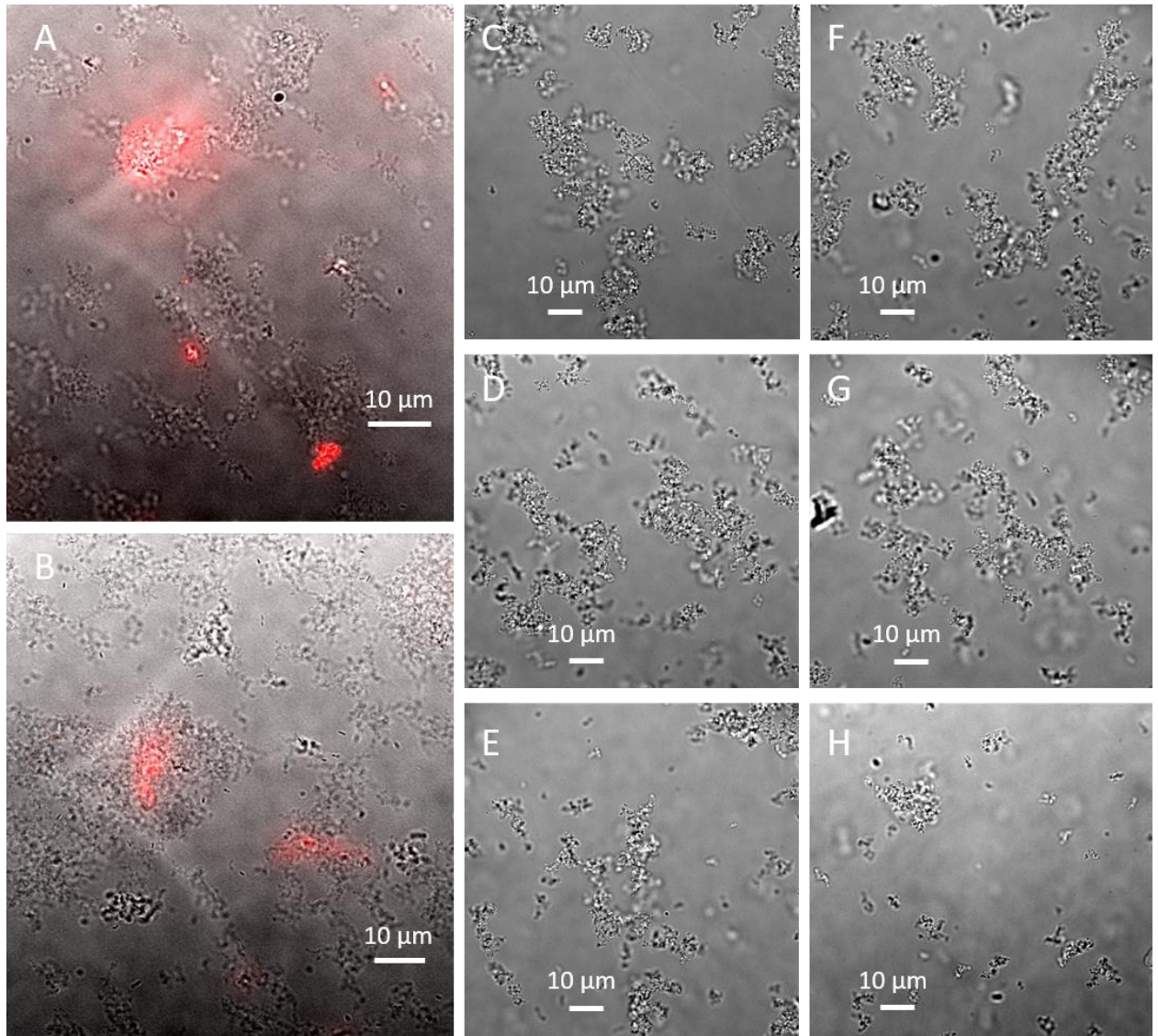
Table 17-8. Supplementary Table 3. Comparison of the different AtzA and AtzC ratio components with their fractal dimensions ( $D_f$ ) and  $\lambda$ .

Protein Components	$D_f (\mu \pm \sigma)$	$\lambda (\mu \pm \sigma)$
A:C (1:8)	$1.60 \pm 0.05$	$0.39 \pm 0.07$
A:C (1:2)	$1.54 \pm 0.01$	$0.48 \pm 0.03$
A:C (3:4)	$1.52 \pm 0.08$	$0.51 \pm 0.15$
A:C (3:2)	$1.66 \pm 0.05$	$0.59 \pm 0.13$
A:C (3:1)	$1.56 \pm 0.09$	$0.61 \pm 0.21$
A:B:C (3:1:2)	$1.49 \pm 0.06$	$0.45 \pm 0.10$
A:B:C (3:1:1)	$1.46 \pm 0.10$	$0.39 \pm 0.04$



**Figure 129-8. Supplementary Figure 33.** DLS and SDS PAGE confirm AtzBSH2 incorporation into the 3-component assembly.

AtzAM1, AtzBSH2, and AtzCM1 were added and allowed to incubate at various concentrations, then analyzed with DLS which showed that the addition of AtzBSH2 continues to have an assembly at ~1  $\mu$ m. The SDS Page gel samples were pelleted and samples of the three component assembly supernatant and pellet were analyzed. If AtzBSH2 is incorporated into the assembly, it should partition preferentially into the pellet. The gels show that a band at the expected MW weight of AtzBSH2 ~69kda is seen predominantly in the pellet with increasing AtzBSH2 concentrations (Also see Fig S34).



**Figure 130-8.** Supplementary Figure 34. Fluorescence microscopy and bright-field images of the 3-component assembly confirm incorporation of AtzBSH2 into assembly while bright-field images confirm the fractal-like nature of the 2-component assembly.

(A and B) 3  $\mu\text{M}$  AtzAM1, 1  $\mu\text{M}$  AtzBSH2 dye labeled with Alexa Fluor™ 647, 2  $\mu\text{M}$  AtzCM1 image shows AtzBSH2 incorporation into 3-component assembly at various locations (C to H) 3  $\mu\text{M}$  AtzAM1 and 2  $\mu\text{M}$  AtzCM1 assembly images depict fractal-like assembly structure.

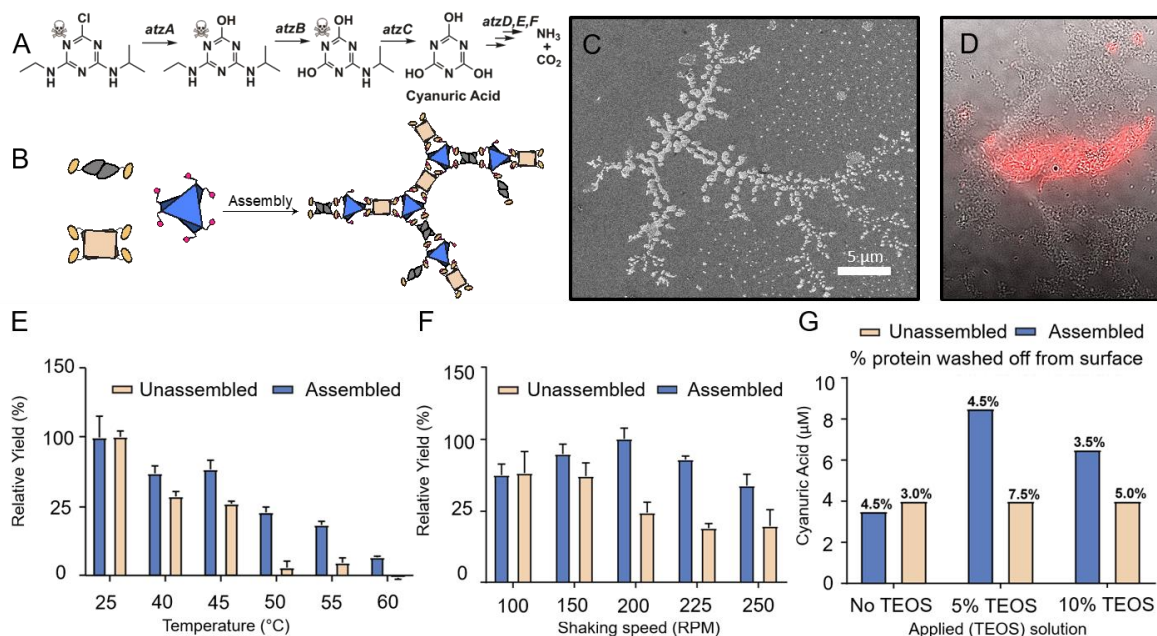
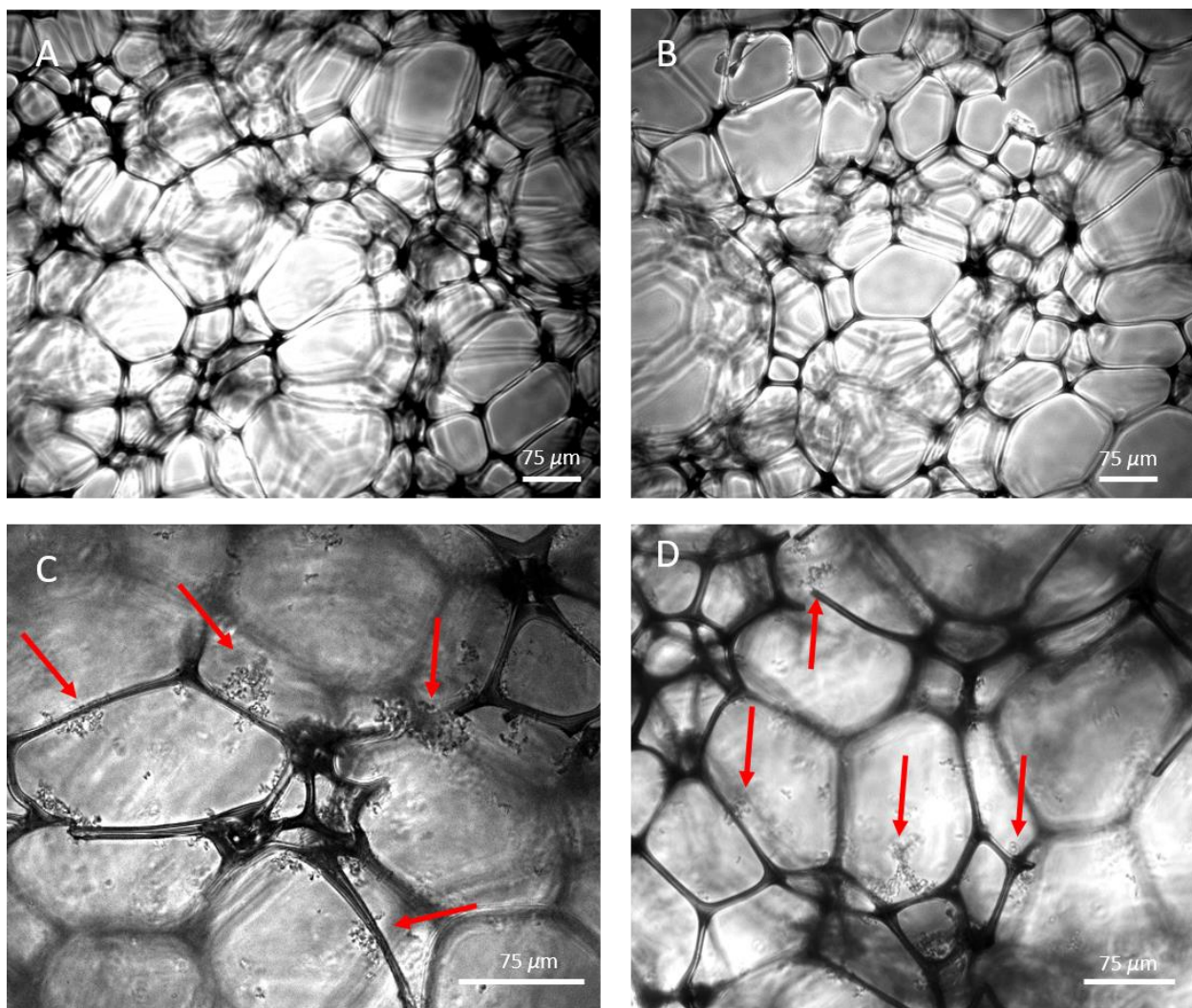


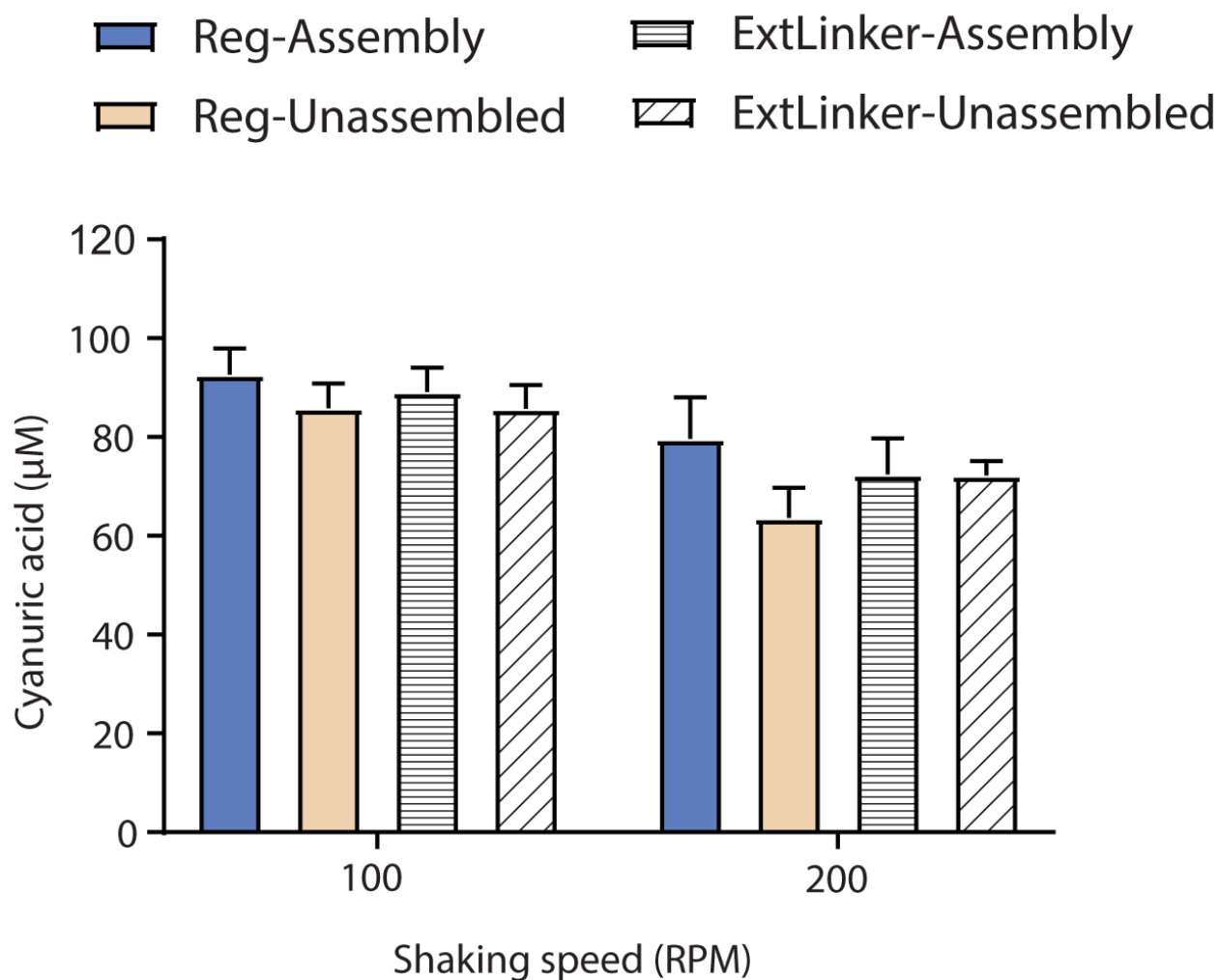
Figure 131-8. Supplementary Figure 35. AtzBSH2 incorporation to construct a three-enzyme assembly.

(A) Atrazine degradation pathway, enzymatic conversion of atrazine to cyanuric acid, and further enzymatic conversion to  $\text{NH}_3$  and  $\text{CO}_2$ . (B) AtzB was added as an SH2-domain fusion to the two-component (AtzA-AtzC) assembly. (C) and (D) Three-component assembly formation was validated using HIM, and the incorporation of AtzB was confirmed with fluorescence microscopy (using an Alexa-658-labeled AtzB). (E) and (F), Assemblies were found to be more thermotolerant, as detected by incubation at a given temperature for 30 min followed by activity assays, and more robust to mechanical shearing forces, as detected by ability to withstand shaking. (G), Assemblies and free enzymes were incorporated into a Basotect® polymer foam with different TEOS % layers, to trap proteins, and assayed for cyanuric acid production. Proteins can be lost during the wash step after crosslinking and the % of protein lost under each condition is indicated on top of the bars.



**Figure 132-8.** Supplementary Figure 36. Phase contrast micrographs of the Basotect® polymer foam with and without assemblies for the AtzAM1, AtzBSH2, and AtzCM1 components.

(A and B) The microporous polymer foam with no assemblies. (C and D) The assemblies have been immobilized into the polymer foam, red arrows depict locations with assemblies. Images were taken with a Leica DM4000 B LED microscope, 10X objective (100X total magnification).



**Figure 133-8.** Supplementary Figure 37. The fractal-like assemblies (Reg-Assembly) and the extended linker globular assemblies (ExtLinker-Assembly) enzymatic conversion of atrazine to cyanuric acid demonstrates no enzymatic benefit of a globular assembly.

(A) AtzB was incorporated into the two-component assembly as an SH2-domain fusion as previously described to create the three-component assembly for both the fractal and globular assemblies. The activity of the fractal assembly was similar than the extended linker assemblies (GS linker) under different shaking speeds.