

Incorporation of non-canonical proline residues into
proteins expressed in *Escherichia coli*

Thesis by
Stephanie L. Breunig

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy in Chemistry

The logo for the California Institute of Technology (Caltech), featuring the word "Caltech" in a bold, orange, sans-serif font.

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
(Defended May 5, 2023)

© 2023

Stephanie L. Breunig
ORCID: 0000-0002-8665-6363

ACKNOWLEDGEMENTS

As with any important transition point in life, writing and defending a doctoral dissertation provides a beautiful opportunity to reflect upon those individuals who have shaped one's trajectory in life. I have been incredibly blessed in the relationships, both professional and personal, that have led me to Caltech and helped me through the trials and tribulations that is graduate school.

Of course, one of the most influential individuals at Caltech has been my advisor, David Tirrell. Incredibly, Dave managed to run an active research group while serving as Caltech's provost for the bulk of my time in his lab. Dave has undoubtedly shaped my scientific thinking beyond that, I think, of a typical advisor. His healthy dose of skepticism, true depth of thought, genuine kindness, and ability to ask *just* the right question are unparalleled. I have grown an incredible amount while a member of his lab, and appreciate the intellectual freedom and independence we were granted as his students.

I gratefully acknowledge the other members of my thesis committee: Dennis Dougherty, Steve Mayo, and Brian Stoltz. I cloned my first construct in Dennis's lab shortly after moving to Pasadena, and appreciate the physical organic chemistry perspective he brought to my education. A collaboration between Steve's group and ours was a fun project that, while not explicitly included in this thesis, refined my thinking around proline conformation and insulin dynamics. Brian taught some of the first classes I took at Caltech, and helped to provide a foundation in organic chemistry. His exceptionally friendly nature (and the occasional pickup volleyball appearance) helped me to first realize that yes, even Caltech professors are people, too.

The members of the Tirrell lab deserve my heartfelt gratitude. The diversity of research interests in the lab has significantly widened my knowledge of so many areas of chemical biology, proteomics, and materials science. Much of this thesis was founded upon work spearheaded by Seth Lieblich and Katharine Fang, and later carried on by Alex Chapman. Their efforts truly made this dissertation possible. To those who have played some role in training and mentoring me—Peter Rapp, Bradley Silverman, Alex Chapman, and Josh Baccile—I cannot thank you enough. You gave me the skills to become the experimentalist

that I am today and provided incredible support and encouragement along the way. The Tirrell lab has changed much over the years, from fifteen or so members when I first joined, to the intrepid five who remain. Elliot Mackrell, Sophie Miller, Hanwei Liu, and Grace Wang: it has been a pleasure to journey through grad school with you, and to see each of you become experts in your subfields. I cannot wait to see how you all go on to change the world.

The volleyball community at Caltech and in Pasadena gave me a competitive outlet, especially early on in my graduate career; it's always satisfying to hit a ball when experiments aren't exactly going your way. The Alpine and Triathlon Clubs at Caltech helped keep me sane in the later years of grad school, often by doing some pretty insane things. They made a road runner fall in love with the trails and all things multisport, and have quite literally changed my life. These folks are some of the most genuine, hard-working, and welcoming (not to mention *fast*) athletes you might ever meet, and I have immensely enjoyed training and racing with them.

I cannot say enough about the Catholic communities around Pasadena. They welcomed me with open arms as soon as I arrived in California (Joanie Noonan, Rachel & Nick Bradley, Mary Washburn), and helped in forming a true Catholic community at Caltech (Jorge & Hayley Llop, Elizabeth Lunny, Tim Csernica, Maria Camarca, Katt Redmond, Lauren Conger, Alessandra Flaherty, Nick Friesenhahn). These friends (and *so* many others) have walked with me in various capacities along my spiritual journey, and have shown me exactly why it's impossible to live the Christian life in isolation.

Some friendships go beyond description. Caitlin Lacker—you are a sister in all but blood. I'm not sure I could have made it through COVID, let alone grad school, without your constant love and support. Thank you for being you.

Finally, I would not be here (in all senses of the phrase) without my family. Mom, Dad, Nicole, Tim, and everyone in my extended family—it's impossible for me to fully appreciate all that you have done for me, let alone explain how grateful I am for it all. Knowing that you have my back, no matter what, has given me the freedom to chase my dreams. There is no greater gift than your love.

ABSTRACT

Non-canonical proline residues expand the chemical space about proline, while maintaining some conformational properties of the canonical residue. The translational machinery of *Escherichia coli* can accommodate close structural analogs of proline, which has enabled the production of recombinant proteins that contain non-canonical residues at proline positions. However, proline mutagenesis in *E. coli* is restricted to a relatively small set of proline variants, and protein science and engineering efforts utilizing non-canonical proline residues are limited.

This thesis aims to expand the scope of proline analogs that can be accepted by *E. coli*, and demonstrate the utility of proline mutagenesis in modifying and studying protein behavior. In Chapter II, we describe the incorporation of three aliphatic proline residues into recombinantly-produced insulin, and find that these modest modifications at ProB28 alter the biophysical properties of the therapeutic protein. In particular, the addition of an exocyclic olefin at B28 accelerated insulin fibril formation, while 4-methyl substituents increased the rate of dissociation from the pharmaceutically-formulated insulin hexamer. We expand our proline mutagenesis approach to monomeric insulins in Chapter III. 4-fluorinated proline analogs replaced ProB29 of the fast-acting insulin lispro; 4S-fluorination of ProB29 slowed fibril formation. Chapter IV describes the incorporation of the photo-activatable proline analog “photo-proline” into proteins expressed in *E. coli*, and Chapter V discusses our efforts to engineer the *E. coli* prolyl-tRNA synthetase to accommodate more diverse proline substrates. Together, this work expands the proline analogs accessible to recombinant expression in *E. coli*, and demonstrates their use in probing and engineering the biophysical properties of proteins.

PUBLISHED CONTENT AND CONTRIBUTIONS

Breunig, S. L. & Tirrell, D. A. “Incorporation of proline analogs into recombinant proteins expressed in *Escherichia coli*.” *Methods Enzymol.* **2021**, *656*, 545-571. doi: 10.1016/bs.mie.2021.05.008.

SLB collected and prepared all data, and participated in the writing of the manuscript.

Excerpts from the introduction, Table 1, and Figures 1 & 2 reprinted with permission from the copyright holder, Elsevier, Inc.

TABLE OF CONTENTS

Acknowledgements	iii
Abstract	v
Published Content and Contributions.....	vi
Table of Contents.....	vii
List of Figures.....	xi
List of Tables	xiii
Nomenclature.....	xiv
Chapter I: Non-canonical proline residues in protein science and engineering. 1	
1.1 Contributions	1
1.2 Abstract.....	1
1.3 Main text.....	2
1.3.1 The conformational properties of proline	3
1.3.2 Non-canonical proline residues.....	3
1.3.3 Proline analogs in modifying protein behavior	7
1.3.4. Incorporation of non-canonical proline residues into polypeptide chains	10
1.4 Overview of thesis chapters	13
1.5 References	15
Chapter II: Incorporation of aliphatic proline residues into recombinantly- produced insulin.....	21
2.1 Contributions	21
2.2 Abstract.....	21
2.3 Introduction	22
2.4 Results and discussion.....	25
2.4.1 Aliphatic proline residues are accepted by the <i>E. coli</i> translational machinery	25
2.4.2 Proline analogs do not affect insulin secondary structure or bioactivity	26
2.4.3 4-methylene at position B28 accelerates fibril formation	28
2.4.4 4-methylation of ProB28 speeds hexamer dissociation	30
2.5 Conclusion.....	32
2.6 Materials and methods	33
2.6.1 Chemicals	33
2.6.2 Enzymes.....	34
2.6.3 Strains and plasmids	35
2.6.4 Primers	36
2.6.5 Nucleotide and amino acid sequences	36
2.6.6 Screening ncPro incorporation.....	38
2.6.7 Proinsulin expression	40
2.6.8 Proinsulin refolding.....	41

2.6.9 Insulin maturation and purification.....	42
2.6.10 MALDI-TOF MS	43
2.6.11 Circular dichroism spectroscopy.....	43
2.6.12 Reduction of blood glucose in diabetic mice.....	45
2.6.13 Fibrillation	45
2.6.14 Transmission electron microscopy	46
2.6.15 ANS fluorescence.....	46
2.6.16 Analytical ultracentrifugation	46
2.6.17 Calculations of proline and proline analog conformation.....	47
2.6.18 Models of ins-4R-Me and ins-4S-Me hexamers.....	47
2.7 Supplementary figures and tables	48
2.8 References	56
Chapter III: 4S-fluorination of ProB29 of insulin lispro slows fibril formation.....	61
3.1 Contributions	61
3.2 Abstract.....	61
3.3 Main text.....	62
3.4 Materials and methods	72
3.4.1 Chemicals	72
3.4.2 Enzymes.....	72
3.4.3 Strains and plasmids	72
3.4.4 Primers	73
3.4.5 Nucleotide and amino acid sequences	73
3.4.6 Proinsulin expression	76
3.4.7 Proinsulin refolding.....	77
3.4.8 Lispro maturation and purification	78
3.4.9 Preparation of lispro variants for mouse experiments.....	79
3.4.10 MALDI-TOF MS	79
3.4.11 Reduction of blood glucose in diabetic mice.....	80
3.4.12 Circular dichroism spectroscopy.....	81
3.4.13 Fibrillation	82
3.4.14 Transmission electron microscopy	83
3.4.15 ANS fluorescence.....	83
3.4.16 Analytical ultracentrifugation	83
3.5 Supplementary figures and tables	84
3.6. References	87
Chapter IV: Incorporation of photo-proline into proteins in <i>Escherichia coli</i>	91
4.1 Contributions	91
4.2 Abstract.....	91
4.3 Introduction	92
4.4 Results and discussion.....	94
4.4.1 Incorporation of photo-proline.....	94
4.4.2 <i>proP</i> overexpression leads to ncPro incorporation.....	96
4.4.3 PPIase substrates were not detected with photo-proline	98
4.5 Conclusion.....	100

4.6	Materials and methods	101
4.6.1	Chemicals	101
4.6.2	Enzymes.....	101
4.6.3	Strains and plasmids	101
4.6.4	Primers	101
4.6.5	Nucleotide and amino acid sequences	102
4.6.6	Cloning.....	104
4.6.7	Incorporation of photo-proline into proinsulin under osmotic stress conditions.....	104
4.6.8	Incorporation of photo-proline into proinsulin with <i>proP</i> overexpression	105
4.6.9	Incorporation of photo-pro with <i>proP</i> overexpression for PPIase crosslinking.....	106
4.6.10	Inclusion body isolation	107
4.6.11	LC-ESI-MS.....	107
4.6.12	UV irradiation.....	108
4.6.13	Western blot.....	109
4.7	Supplementary figures and tables	110
4.8	References	115
Chapter V: Toward expanding non-canonical proline incorporation by prolyl-tRNA synthetase engineering.....		118
5.1	Contributions	118
5.2	Abstract.....	118
5.3	Introduction	119
5.4	Results and discussion.....	121
5.4.1	Full-length fluorescent proteins are poor reporters of proline analog incorporation	121
5.4.2	Split-GFP as a reporter for proline analog incorporation.....	123
5.4.3	Efforts to improve split-GFP dynamic range reveals poor GFP1-10 solubility	126
5.4.4	Solubility tags improve GFP1-10 solubility	129
5.4.5	Changing expression protocols does not improve screen performance	133
5.4.6	Alternative GFP11 designs do not improve performance.....	135
5.4.7	Inclusion body analysis by flow cytometry	137
5.4.8	Survival on proline analog-containing medium	138
5.4.9	A small molecule alternative to fluorescent reporter proteins.....	138
5.4.10	Optimization of FIAsh labeling conditions does not improve sorting performance.....	141
5.4.11	Considerations for future ProRS engineering.....	147
5.5	Conclusion	150
5.6	Materials and methods	151
5.6.1	Chemicals	151
5.6.2	Enzymes.....	152

5.6.3	Strains	152
5.6.4	DNA oligos	153
5.6.5	g-Block gene fragments.....	156
5.6.6	Plasmids and cloning.....	158
5.6.7	Library generation	163
5.6.8	Nucleotide and amino acid sequences	165
5.6.9	Non-canonical proline incorporation and protein expression	170
5.6.10	Quantification of ncPro incorporation	171
5.6.11	Western blot analysis of protein solubility	172
5.6.12	FLAsH labeling (optimized conditions).....	173
5.6.13	Flow cytometry and FACS.....	173
5.6.14	Replica plating.....	174
5.6.15	Identification of proline-free carrier proteins	174
5.6.16	Microscopy	175
5.7	Supplementary figures and tables	176
5.8	References	186

LIST OF FIGURES

Figure 1.1. Proline conformational preferences.....	3
Figure 1.2. Non-canonical amino acid mutagenesis.....	11
Figure 2.1. Proline mutagenesis at position B28 of human insulin.....	23
Figure 2.2. Mass spectrometry characterization of insulin variants.....	26
Figure 2.3. Circular dichroism spectroscopy, sedimentation, and bioactivity of insulin variants	27
Figure 2.4. Fibrillation of insulin variants	29
Figure 2.5. Hexamer dissociation kinetics of insulin variants.....	31
Figure 2.6. Sedimentation velocity of insulin and ins-4 <i>S</i> -Me	32
Figure 2.S1. SDS-PAGE after non-canonical proline incorporation	48
Figure 2.S2. Purity of insulin samples assessed by SDS-PAGE.....	49
Figure 2.S3. Circular dichroism controls	50
Figure 2.S4. Models of ins-4 <i>R</i> -Me and ins-4 <i>S</i> -Me hexamers	51
Figure 2.S5. Conformations of proline, 4-methyleneproline, and 3,4-dehydroproline.....	52
Figure 3.1. Proline mutagenesis at position B29 of insulin lispro	63
Figure 3.2. Mass spectrometry characterization of lispro variants.....	67
Figure 3.3. Bioactivity, circular dichroism spectroscopy, and sedimentation of lispro variants.....	68
Figure 3.4. Hexamer dissociation kinetics of lispro variants	69
Figure 3.5. Fibrillation of lispro variants	70
Figure 3.S1. Purity of lispro samples assessed by SDS-PAGE	84
Figure 3.S2. Circular dichroism controls	85
Figure 4.1. Improved incorporation of photo-pro	95
Figure 4.2. <i>proP</i> overexpression leads to proline analog incorporation	97
Figure 4.3. PPIase:client interactions were not detected using photo-proline	99
Figure 4.S1. Thermal degradation products of photo-pro include dhp.....	110
Figure 5.1. mWasabi fluorescence does not correspond to proline analog incorporation efficiencies	122

Figure 5.2. Design of split-GFP to measure proline analog incorporation	124
Figure 5.3. 4S-NH ₂ incorporation does not improve after sorting	125
Figure 5.4. Leaky GFP11 expression and residual proline result in high background fluorescence	127
Figure 5.5. Poor solubility impairs split-GFP performance	127
Figure 5.6. Staggered expression of split-GFP at 25°C.....	130
Figure 5.7. Fluorescence of split supercharged and permuted GFP variants.....	131
Figure 5.8. GFP1-10 solubility tag fusions improve resolution by flow cytometry	132
Figure 5.9. GFP1-10/11 reverse expression.....	134
Figure 5.10. SPI expression of GFP1-10/11	134
Figure 5.11. XTEN-GFP11 fusion proteins to measure proline analog incorporation.....	136
Figure 5.12. Interrogating inclusion body formation by flow cytometry	137
Figure 5.13. A FAsH-based labeling strategy to measure proline analog incorporation.....	139
Figure 5.14. Optimization of FAsH labeling and protein expression conditions .	141
Figure 5.15. Sorting after FAsH labeling reveals poor rescue efficiency and Top7 toxicity.....	142
Figure 5.16. Alternative proline-free target proteins	144
Figure 5.17. Validation of the DHR14 approach.....	145
Figure 5.18. Significant problems persist	147
Figure 5.S1. Proline-containing TC motifs do not accurately measure proline analog incorporation	176
Figure 5.S2. Development of a proline-free TC motif	177
Figure 5.S3. FAsH labeling after Top7 expression leads to heterogeneous fluorescence	178
Figure 5.S4. FAsH labeling after DHR14-TC3 expression can measure proline analog incorporation	179
Figure 5.S5. ProRS overexpression increases proline analog incorporation	180
Figure 5.S6. Cheaters overwhelm library samples after sorting	181
Figure 5.S7. Plasmid stability over rounds of transformation and growth	182
Figure 5.S8. Growth depends upon ProRS variant and promoter.....	183

LIST OF TABLES

Table 1.1. Proline analogs used for residue-specific incorporation into recombinant proteins	5
Table 2.S1. Incorporation of non-canonical proline residues.....	53
Table 2.S2. Mass spectrometry characterization of insulin variants.....	54
Table 2.S3. Expression conditions and insulin yields	54
Table 2.S4. Summary of insulin variant characterization	54
Table 2.S5. Sedimentation coefficients of insulin and ins-4S-Me	55
Table 3.S1. Mass spectrometry characterization of lispro variants.....	86
Table 3.S2. Lispro yields.....	86
Table 3.S3. Summary of lispro variant characterization	86
Table 4.1. Photo-pro incorporation efficiencies under osmotic stress conditions	96
Table 4.2. Proline analog incorporation efficiencies with <i>proP</i> overexpression	97
Table 4.S1. 3x-FLAG-tagged PPIases	113
Table 5.S1. Mutation rates for error-prone PCR library generation	184
Table 5.S2. Amino acid sequences of proline-free proteins.....	185

NOMENCLATURE

- 2-Me.** 2-methylproline (or α -methylproline)
3R-OH. 3R-hydroxyproline
3S-OH. 3S-hydroxyproline
44-diF. 4,4-difluoroproline
44-diMe. 4,4-dimethylproline
4ene. 4-methyleneproline
4-keto. 4-oxoproline
4R-F. 4R-fluoroproline
4R-Me. 4R-methylproline
4S-F. 4S-fluoroproline
4S-Me. 4S-methylproline
4S-NH₂. 4S-aminoproline
4SSM. 4-site saturation mutagenesis
AA. Amino acid
aaRS. Aminoacyl-tRNA synthetase
Aha. Azidohomoalanine
Anl. Azidonorleucine
ANS. 8-anilino-1-naphthalenesulfonic acid
AUC. Analytical ultracentrifugation
Aze. Azetidine-2-carboxylic acid
 β 2m. β 2 microglobulin
BME. β -mercaptoethanol
BONCAT. Bioorthogonal non-canonical amino acid tagging
CD. Circular dichroism
CFUs. Colony-forming units
CSII. Continuous subcutaneous insulin infusion
dhp. 3,4-dehydroproline
EDT. Ethanedithiol

EGFP. Enhanced green fluorescent protein

ePCR. Error-prone PCR

FACS. Fluorescence-activated cell sorting

FAI. Fast-acting insulin

FLAsH-EDT₂. fluorescein arsenical hairpin binder-ethanedithiol

GFP. Green fluorescent protein

HPLC. High performance liquid chromatography

IPTG. Isopropyl β -D-1-thiogalactopyranoside

KP. Insulin lispro

LB. Luria Bertani (medium)

MALDI-TOF MS. Matrix-assisted laser desorption/ionization-time of flight mass spectrometry

MetRS. Methionyl-tRNA synthetase

mRFP1. Monomeric red fluorescent protein 1

ncAA. Non-canonical amino acid

ncPro. Non-canonical proline

PDB. Protein databank

Pip. Piperidine-2-carboxylic acid

Pip-Az. Piperazine-2-carboxylic acid

Pip-OH. (2*S*, 5*S*)-5-hydroxypiperidine-2-carboxylic acid

PPIase: Peptidyl-prolyl isomerase

ProRS. Prolyl-tRNA synthetase

TC. Tetracysteine

TFA. Trifluoroacetic acid

Tfn. Trifluoronorleucine

ThT. Thioflavin T

Thz. 1,3-thiazoline-4-carboxylic acid

Trx. Thioredoxin

*Chapter I*NON-CANONICAL PROLINE RESIDUES IN PROTEIN SCIENCE AND
ENGINEERING**1.1 Contributions**

Stephanie L. Breunig and David A. Tirrell contributed to the writing of this text.

1.2 Abstract

Proline residues are unique in the extent to which they constrain the conformational space available to the protein backbone. Because the conformational preferences of proline cannot be recapitulated by any of the other proteinogenic amino acids, standard mutagenesis approaches that seek to introduce new chemical functionality at proline positions unavoidably perturb backbone flexibility. Here, we describe some unique conformational properties of proline and non-canonical proline variants, and their use in modifying and understanding protein behavior. We further detail the incorporation of proline analogs into recombinant proteins in *Escherichia coli* via a residue-specific mutagenesis strategy. The ability of proline analogs to maintain backbone conformational restraints while introducing new chemistries grant them a useful and intriguing role in protein science and engineering endeavors.

1.3 Main text

1.3.1 The conformational properties of proline

As the sole α -imino acid among the canonical building blocks of proteins, proline is unique. Its pyrrolidine ring imposes conformational constraints on the polypeptide backbone that are essential to protein structure and function (Figure 1.1a,b). The backbone dihedral angle ϕ of proline residues is restricted to $63\pm 15^\circ$ (Ref. 1), limiting the conformational trajectories that are available to the polypeptide chain. The reduced conformational space sampled by the peptide backbones of proline residues is illustrated by the Ramachandran plots for proline and the acyclic amino acids (Figure 1.1a,b). The pyrrolidine side chain is also limited in its conformational flexibility, and generally adopts one of only two rapidly-interconverting ring puckers: C^γ -*endo* or C^γ -*exo*² (Figure 1.1c-d). The conformation of the five-membered ring can be described quantitatively through the concept of pseudorotation, in which the two defining parameters, phase angle and amplitude, are calculated from the pyrrolidine torsional angles.^{3,4} The highly populated phase angles of the C^γ -*endo* and -*exo* puckers are approximately 198° and 18° , respectively; intermediate phase angles describe other higher energy envelope and twist conformations.

Because the proline amide linkage cannot serve as a hydrogen-bond donor, the presence of proline residues within α -helical or β -strand structural motifs is disfavored.⁵ Proline's *cis* and *trans* conformers are nearly isoenergetic, and the barrier to *cis-trans* isomerization is reduced in comparison to the other amino acids. Consequently, *cis* isomers are more common for proline than for any other canonical amino acid,¹ and *cis-trans* isomerization at proline residues can play important roles in protein folding⁶ and function.⁷

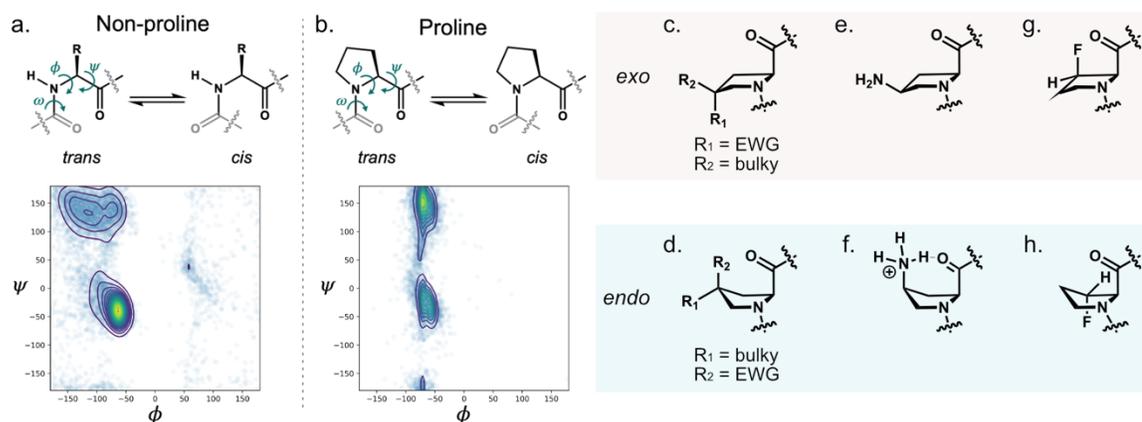


Figure 1.1. Proline conformational preferences. **a-b.** Backbone conformations of, and Ramachandran plots for, non-proline (**a**) and proline (**b**) residues found in structural data deposited in the Protein Data Bank (PDB). The torsional angles ϕ , ϕ' , and ω , along with *cis* and *trans* amide isomers are indicated in the amino acid structures. **c-h.** *Exo* and *endo* ring pucker preferences for select proline analogs. ncPro residues with 4*R*- electron-withdrawing groups or bulky 4*S*- substituents prefer the *exo* pucker (**c**); those with 4*S*- electron-withdrawing groups or bulky 4*R*- substituents prefer *endo* (**d**). In basic conditions, 4*S*-NH₂ prefers the pseudoequatorial position in the *exo* ring pucker (**e**). When protonated, the ammonium group functions as an electron-withdrawing substituent; a transannular hydrogen bond further stabilizes the *endo* pucker (**f**). 3*R*-F (**g**) and 3*S*-F (**h**) prefer the *exo* and *endo* puckers, respectively, due to a gauche effect.

1.3.2 Non-canonical proline residues

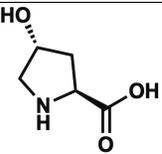
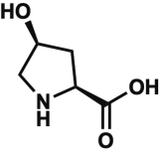
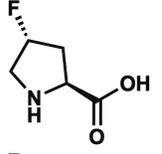
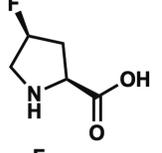
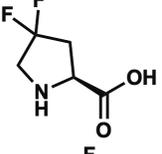
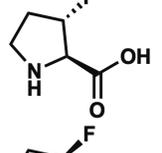
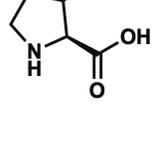
Non-canonical amino acids (ncAAs) have found extensive use in chemical biology and related fields.⁸⁻¹⁰ They serve many roles, such as providing chemical handles for protein modification,¹¹ serving as probes in time-resolved and cell-selective proteomic analyses,¹² interrogating the effects of post-translational modifications,¹³ identifying protein-protein interaction partners,¹⁴ tracking protein location in vivo,¹⁵ and probing the importance of non-covalent interactions in protein behavior.¹⁶

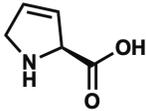
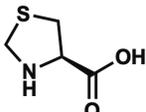
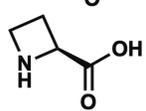
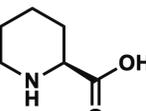
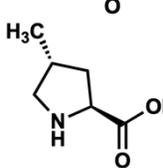
The utility of standard mutagenesis at proline sites is limited by its impact on chain conformation. ncAAs provide a means of addressing this limitation, as replacement of proline by non-canonical analogs allows introduction of new chemical functionality while maintaining conformational constraints. For instance, the hydroxyl groups of the non-canonical proline (ncPro) variants 4*R*-hydroxyproline (4*R*-OH) and 4*S*-hydroxyproline (4*S*-OH, Table 1.1) permit hydrogen-bonding interactions; one such hydrogen bond has been suggested to alter the behavior of an engineered insulin variant.¹⁷ Expanding or contracting the five-membered pyrrolidine ring can add (piperidine-2-carboxylic acid, Pip, Table 1.1) or remove (azetidine-2-carboxylic acid, Aze, Table 1.1) hydrophobic packing interactions.¹⁸ The alkene functionality in 3,4-dehydroproline (Dhp, Table 1.1) has been used as a chemical handle to modify protein-based materials.¹⁹ Fluorinated proline variants can be used as conformational reporters in ¹⁹F NMR experiments.²⁰ For instance, *cis-trans* isomerization of the sole 4,4-difluoroproline (44-diF, Table 1.1) residue in a modified β2 microglobulin could be monitored by ¹⁹F NMR.²¹

Many ncPro analogs have well-documented conformational biases; we here describe the conformations of a selection of proline analogs, though others are extensively reviewed elsewhere.^{2,22} Most widely studied are proline variants containing substituents at the C^γ position. The presence of a 4*R*- electron-withdrawing substituent (as in 4*R*-OH and 4*R*-fluoroproline, 4*R*-F; Table 1) stabilizes the C^γ-*exo* ring pucker through a gauche effect (Figure 1.1c). The *exo* ring pucker, in turn, pre-organizes the amide in the *trans* conformation. Conversely, ncPro residues with 4*S*- electron-withdrawing groups (such as 4*S*-OH and 4*S*-fluoroproline, 4*S*-F; Table 1.1) favor the *endo* ring pucker (Figure 1.1d),

and have a higher propensity for the *cis* amide isomer compared to canonical proline. The presence of a 4-fluoro substituent (and presumably any electron-withdrawing group) lowers the energy barrier to *cis-trans* isomerization, as the substituent decreases the bond order of the preceding amide.²³

Table 1.1: Proline analogs used for residue-specific incorporation into recombinant proteins

Proline analog	Proline structure	ProRS variant	Reported [NaCl] (M)	Reported level of proline replacement	Reference
4 <i>R</i> -OH		wt	0.3	88%	17
4 <i>S</i> -OH		wt	0.3	91%	17
4 <i>R</i> -F		wt	0.3	96%	24
4 <i>S</i> -F		wt	0.3	97%	24
4,4-diF		wt	0.3	86%	24
3 <i>S</i> -F		—	—	“virtually complete”	25
3 <i>R</i> -F		—	—	“virtually complete”	25

Proline analog	Proline structure	ProRS variant	Reported [NaCl] (M)	Reported level of proline replacement	Reference
dhp		wt	0.3	≤100%	18
Thz		wt	0.3	90%	18
Aze		wt	0.3	≤100%	18
Pip		C443G	0.3	89%	18
4 <i>R</i> -Me		C443G	0.6	>60%	26

Sterically-demanding groups at the C^{γ} position instead prefer the pseudoequatorial position, rather than pseudoaxial favored by electron-withdrawing substituents. 4*R*-methylproline (4*R*-Me) thus enforces the *endo* ring pucker and *cis* amide isomer compared to proline (Figure 1.1d), while the 4*S* diastereomer (4*S*-Me) prefers the *exo* pucker and *trans* amide (Figure 1.1c).²⁷ Similar effects were observed for the 4-mercaptoproline derivatives,²⁸ though their conformational biases are less pronounced. Other bulky substituents at the C^{γ} position are expected to behave similarly.

One especially interesting 4-substituted proline analog is 4*S*-aminoproline (4*S*-NH₂).²⁹ Changes in amine protonation state over physiologically-relevant pH ranges alter the ring puckering preferences of this analog: under basic conditions, sterics dominate, and the *exo* conformation is preferred (Figure 1.1e). The ammonium formed upon protonation behaves

as an electron-withdrawing substituent, and a transannular hydrogen bond with the backbone carbonyl further reinforces the preference for the *endo* pucker (Figure 1.1f). In the case of 4*S*-NH₂, both protonation states prefer the *trans* amide isomer. These properties of 4*S*-NH₂ have led to pH-responsive collagen-like peptides.²⁹

Functionalization at the C^β position, as in 3*R*-F and 3*S*-F, similarly leads to conformational biases relative to proline. 3*R*-F prefers the C^γ-*exo* pucker (Figure 1.1g), and 3*S*-F the *endo* (Figure 1.1h).²⁵ The conformational preferences of 3*R*-F and 3*S*-F prolines are opposed to those of prolines with 4-fluoro substituents (Figure 1.1c-d). In the context of a polypeptide chain, α-methylproline (2-Me) exists nearly exclusively as the *trans* amide isomer, due to steric clash between the α-substituent and the side chain of the preceding residue.³⁰ The backbone trajectories of 2-Me are even more constrained than that of proline: this residue overwhelmingly prefers the 3₁₀/α-helical conformation and promotes β-turns at the *i* + 1 corner position.³¹

1.3.3 Proline analogs in modifying protein behavior

The established conformational preferences of proline analogs provide a means of interrogating the effects of conformational properties (such as pyrrolidine ring pucker or amide isomerization) on protein structure and function.²² For instance, ncPro residues have been used to identify a key proline *cis-trans* isomerization event in ion channel opening,⁷ demonstrate the stereoelectronic basis of collagen stability,² and modulate the properties of protein-based materials.³²

In many cases in which the conformational preference of the ncPro residue matches that in the native protein structure, the protein fold is stabilized.^{33–36} For instance, in the first example of the biosynthetic incorporation of fluorinated proline residues, barstar Pro46 (*cis* in the native protein) was replaced with 4*R*-F, 4*S*-F, and 4,4-diF.³⁶ Pro46 exists in the *cis* conformation in the native protein. Satisfyingly, the stability of the fluorinated barstar variants correlated well with the *cis-trans* isomerization preferences of the individual proline analogs.

However, protein behavior is complex, and not all reported examples are straightforward.^{37–41} In its native fold, thioredoxin (Trx) contains a proline residue in the *cis* conformation; *trans* to *cis* isomerization is the rate-limiting step in Trx folding, and is essential for Trx function. Despite the opposed *cis-trans* preferences of 4*R*-F and 4*S*-F, proline monofluorination destabilized Trx independent of stereochemistry.³⁹ 4,4-difluorination, which should lower the energy barrier to *cis-trans* isomerization,³⁶ did not accelerate Trx folding.⁴¹ In another context, the two proline residues in the foldon domain were replaced with six proline analogs carrying C γ substituents; changes in stability did not correlate with known *cis-trans* isomerization preferences.³⁸

β 2 microglobulin (β 2m) fibrillation was probed by proline mutagenesis in an example that exemplifies the complexity of protein biophysics. Pro32, which exists in the *cis* conformation in the native state, but *trans* in the fibril, was replaced with 4*R*-F, 4*S*-F, and 4,4-diF.²¹ Stability against chemical denaturation correlated well with the established *cis-trans* isomerization preferences of the proline analogs, suggesting that the backbone

conformation at Pro32 is important for β 2m behavior. Interestingly, 44-diF, which exhibits similar *cis-trans* isomer preferences as proline, led to increased structural disorder and the greatest propensity to form amorphous aggregates. This effect was attributed to a lowered energy barrier to *cis-trans* isomerization, rather than a thermodynamic preference for one isomer. The authors observed diverse fibril morphologies and fibrillation behaviors after seeding among the fluorinated β 2m variants.²¹

In a follow-up study, Pro32 was replaced with 2-Me.⁴² This residue's enhanced preference for the *trans* isomer was hypothesized to promote β 2m fibril formation. While α -methylation of β 2m did lead to greater conformational flexibility of the protein and promoted oligomerization, fibril formation was not observed unless the sample was seeded with pre-formed fibrils. Together, these studies demonstrate that β 2m fibrillation is more complex than simple association of *trans*-Pro32 β 2m conformers, and highlight the unexpected effects that functionalized proline variants can have on protein behavior.^{21,42}

The effect of proline analogs becomes more complex when many proline residues are modified. Global proline mutagenesis has accelerated the folding of EGFP⁴³ and mRFP1,⁴⁴ and stabilized a single-chain Fv fragment.⁴⁵ Simultaneous incorporation of 4S-F (replacing six proline residues), 4-fluoro-phenylalanine (sixteen phenylalanine residues) and 6-fluorotryptophan (2 tryptophan residues) into lipase did not significantly affect the protein's structure or catalytic activity.⁴⁶ Incredibly, replacing 32 proline residues in the KlenTaq DNA polymerase with 4R-F led to an active polymerase with no appreciable differences in activity,⁴⁷ though thermal stability was modestly affected.⁴⁸ In many of these

examples, incorporation of the opposite ncPro diastereomer led to insolubility^{43,44} or apparent instability^{26,47} during recombinant expression. Due to the number of proline residues modified in these cases, it is difficult to attribute changes in protein behavior to any single effect.

1.3.4 Incorporation of non-canonical proline residues into polypeptide chains

Virtually any ncAA can be introduced into a polypeptide chain through chemical means (i.e., solid phase peptide synthesis). However, chemical synthesis approaches are typically limited by polypeptide length and scalability, and restrict the ability to use ncAAs *in vivo*. Alternatively, ncAAs can be introduced into ribosomally-synthesized proteins by hijacking a cell's translational machinery. Two general strategies have been developed for *in vivo* introduction of ncAAs into proteins: site-specific and residue-specific replacement. Both approaches rely on the availability of aminoacyl-tRNA synthetases (aaRSs) able to charge their cognate tRNAs with the ncAAs of interest (Figure 1.2a). Site-specific approaches (including nonsense suppression and related techniques) cause minimal perturbation of protein sequence, and important advances in such methods (for example, Ref. 49–51) have been accomplished since the first report of incorporation in *Escherichia coli*.⁵² In site-specific ncAA mutagenesis, a re-assigned codon (most often the amber stop codon) is matched with the ncAA of interest by an engineered, orthogonal tRNA/aaRS pair. The resulting ncAA-charged tRNA competes with release factors for amber codon recognition within the ribosomal complex, and successful translational read-through positions the ncAA at the amber site.⁸ Challenges in implementing site-specific methods include the formation of truncation products that reduce recombinant protein yields, especially when

ncAAs are incorporated at multiple positions; and the requirement for development of orthogonal tRNA/aaRS pairs, which can be difficult.

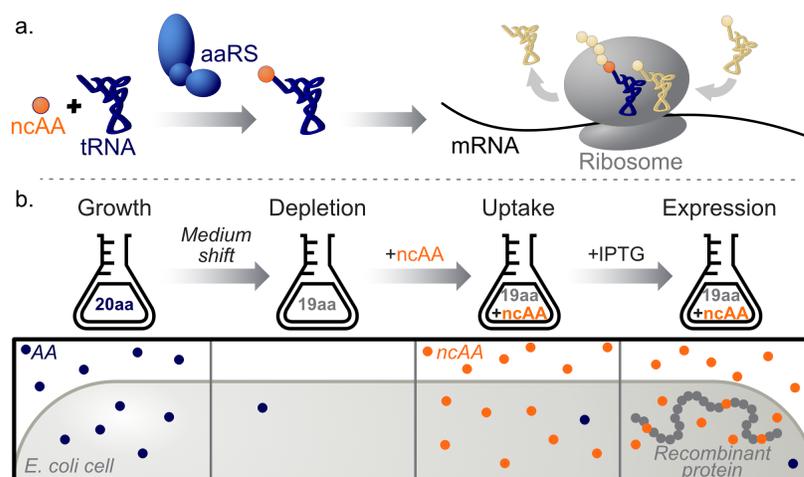


Figure 1.2. Non-canonical amino acid (ncAA) mutagenesis. **a.** For both site-specific and residue-specific approaches, incorporation of ncAA residues into recombinant proteins relies on the ability of the translational machinery of the host to accommodate the ncAA of interest. In many cases, the limiting step is aminoacylation by the relevant aminoacyl-tRNA synthetase (aaRS, left). Upon aminoacylation, the charged tRNA is delivered to the ribosome, which adds the ncAA to the growing polypeptide chain (right). **b.** Residue-specific ncAA mutagenesis workflow. After growth in medium that contains the canonical amino acid (AA) to be replaced, a shift to AA-depleted medium is performed. After addition of the ncAA of interest, expression of the recombinant protein of interest is induced. The ncAA replaces the canonical AA residues in all newly synthesized proteins, including the recombinant protein.

In contrast to site-specific methods, residue-specific ncAA mutagenesis⁹ results in the global replacement of a canonical amino acid with a non-canonical counterpart. This technique typically uses defined expression media and amino acid auxotrophs (i.e., strains deficient in biosynthesis of the amino acid of interest) as expression hosts. A generalized workflow for residue-specific ncAA mutagenesis is depicted in Figure 1.2b. In some cases,

the endogenous aaRS is promiscuous enough to allow the desired substitution;¹² in other cases, overexpression of the endogenous aaRS⁵³ or expression of a mutant aaRS is required.⁵⁴ Residue-specific replacement often requires less genetic manipulation than site-specific techniques, and can produce high yields of the recombinant protein of interest: typical yields are 50-60% of those obtained from expression in media that contain the canonical amino acid.

Residue-specific ncAA mutagenesis has been used in a variety of contexts. Global incorporation of the methionine analog azidohomoalanine (Aha) enables ‘bio-orthogonal non-canonical amino acid tagging’ (BONCAT), a proteomic method capable of enriching for newly synthesized proteins.¹² Various methionine analogs have also been incorporated by residue-specific methods to study purified proteins, including elastin-based biomaterials⁵⁵ and prion proteins.⁵⁶ Leucine analogs have been used to alter the properties of coiled-coil proteins⁵⁷ and introduced into GFP.⁵⁸

Global incorporation of proline analogs has been used to modulate the properties of purified, recombinant proteins. Many of the examples of proline mutagenesis in full-length proteins discussed above were generated by recombinant expression. The first example of biosynthetic ncPro incorporation in recombinant proteins expressed in *E. coli* was the replacement of five proline residues in human annexin V with 1,3-thiazolidine-4-carboxylic acid (Thz).⁵⁹ Since then, additional proline variants have been reported to show translational activity in *E. coli*.^{23,60} The most commonly used analogs are 4R-F and 4S-F, which have been found to alter the folding and fluorescence properties of fluorescent

proteins^{43,44} the melting temperatures of elastin-mimetic peptides.^{25,32} Similar approaches have been used to incorporate proline analogs into recombinantly-expressed barstar,²³ ubiquitin,³⁵ thioredoxin,^{26,39} DNA polymerase,⁴⁷ and 4-oxalocrotonate tautomerase.⁶¹ We have found that replacement of proline B28 of human insulin by a variety of proline analogs can be used to engineer the therapeutically-relevant biophysical properties of the protein.^{17,18} Because no aaRS/tRNA pair capable of site-specific incorporation of proline analogs has been described to date, residue-specific ncAA mutagenesis provides the only option currently available for introducing modified proline residues into recombinant proteins produced in living cells.

The *E. coli* prolyl-tRNA synthetase (ProRS) and downstream translational machinery have been shown to accommodate a variety of structural analogs of proline. For these analogs, simple overexpression of the *E. coli* ProRS (or point mutants thereof), combined with increased expression of proline transporters under hyperosmotic conditions,⁶² enables high levels of proline replacement.⁶⁰ Proline analogs that have been incorporated into recombinant proteins are detailed in Table 1.1, along with reported expression conditions.

1.4 Overview of thesis chapters

In this thesis, I discuss the residue-specific incorporation of proline analogs into proteins expressed in *E. coli*. Chapters II and III build upon previous work in our lab^{17,18,24,63} to engineer the therapeutic protein insulin by proline mutagenesis. In Chapter II, we replace ProB28 of human insulin with three aliphatic proline analogs and evaluate their effect on insulin biophysics; the biosynthetic incorporation of two of these analogs had not

previously been reported. In Chapter III, we expand these proline mutagenesis efforts to the fast-acting insulin lispro, replacing ProB29 with a set of 4-fluorinated proline residues. These two chapters demonstrate the utility of ncAA mutagenesis in probing and modifying the therapeutically-relevant properties of protein drugs.

Chapter IV characterizes the incorporation of photo-proline, a photo-activatable proline analog, into proteins in *E. coli*. We also describe initial attempts in studying the substrates of peptidyl-prolyl isomerases in *E. coli* using this residue. Finally, proline analog incorporation has thus far been fairly limited with respect to the diversity of proline analogs accepted by *E. coli*. In Chapter V, we describe our efforts to engineer the *EcProRS* to recognize an expanded set of ncPro residues. Together, this work expands the number of proline analogs incorporated into recombinant proteins, and demonstrates their utility in studying and engineering protein behavior.

1.5 References

- (1) MacArthur, M. W.; Thornton, J. M. Influence of Proline Residues on Protein Conformation. *J Mol Biol* **1991**, *218*, 397–412.
- (2) Shoulders, M. D.; Raines, R. T. Collagen Structure and Stability. *Annu Rev Biochem* **2009**, *78*, 929–958. <https://doi.org/10.1146/annurev.biochem.77.032207.120833>.
- (3) Park, H. S.; Kang, Y. K. Puckering Transition of the Proline Residue along the Pseudorotational Path: Revisited. *New J Chem* **2021**, *45* (22), 9780–9793. <https://doi.org/10.1039/D1NJ01361K>.
- (4) DeTar, D. L. F.; Luthra, N. P. Conformations of Proline. *J Am Chem Soc* **1977**, *99* (4), 1232–1244. <https://doi.org/10.1021/ja00446a040>.
- (5) Levitt, M. Conformational Preferences of Amino Acids in Globular Proteins. *Biochemistry* **1978**, *17* (20), 4277–4285. <https://doi.org/10.1021/BI00613A026>.
- (6) Lu, K. P.; Finn, G.; Lee, T. H.; Nicholson, L. K. Prolyl *Cis-Trans* Isomerization as a Molecular Timer. *Nat Chem Biol* **2007**, *3* (10), 619–629. <https://doi.org/10.1038/nchembio.2007.35>.
- (7) Lummis, S. C. R.; Beene, D. L.; Lee, L. W.; Lester, H. A.; Broadhurst, R. W.; Dougherty, D. A. *Cis-Trans* Isomerization at a Proline Opens the Pore of a Neurotransmitter-Gated Ion Channel. *Nature* **2005**, *438*, 248–252. <https://doi.org/10.1038/nature04130>.
- (8) Chin, J. W. Expanding and Reprogramming the Genetic Code. *Nature* **2017**, *550*, 53–60. <https://doi.org/10.1038/nature24031>.
- (9) Johnson, J. A.; Lu, Y. Y.; Van Deventer, J. A.; Tirrell, D. A. Residue-Specific Incorporation of Non-Canonical Amino Acids into Proteins: Recent Developments and Applications. *Curr Opin Chem Biol* **2010**, *14*, 774–780. <https://doi.org/10.1016/j.cbpa.2010.09.013>.
- (10) Mukai, T.; Lajoie, M. J.; Englert, M.; Soll, D. Rewriting the Genetic Code. *Annu Rev Microbiol* **2017**, *71*, 557–577. <https://doi.org/10.1146/annurev-micro-090816>.
- (11) Link, A. J.; Vink, M. K. S.; Tirrell, D. A. Presentation and Detection of Azide Functionality in Bacterial Cell Surface Proteins. *J Am Chem Soc* **2004**, *126*, 10598–10602. <https://doi.org/10.1021/ja047629c>.
- (12) Dieterich, D. C.; Link, A. J.; Graumann, J.; Tirrell, D. A.; Schuman, E. M. Selective Identification of Newly Synthesized Proteins in Mammalian Cells Using Bioorthogonal Noncanonical Amino Acid Tagging (BONCAT). *Proc Natl Acad Sci* **2006**, *103* (25), 9482–9487. <https://doi.org/10.1073/pnas.0601637103>.
- (13) Luo, X.; Ju, G.; Wang, R. E.; Zhu, X.; Zambaldo, C.; Liu, R.; Liu, T.; Lyu, X.; Du, J.; Xuan, W.; Yao, A.; Reed, S. A.; Kang, M.; Zhang, Y.; Guo, H.; Huang, C.; Yang, P.-Y.; Wilson, I. A.; Schultz, P. G.; Wang, F. Genetically Encoding Phosphotyrosine and Its Nonhydrolyzable Analog in Bacteria. *Nat Chem Biol* **2017**, *13*, 845–849. <https://doi.org/10.1038/nchembio.2405>.

- (14) Chin, J. W.; Martin, A. B.; King, D. S.; Wang, L.; Schultz, P. G. Addition of a Photocrosslinking Amino Acid to the Genetic Code of *Escherichia coli*. *Proc Natl Acad Sci* **2002**, *99* (17), 11020–11024. <https://doi.org/10.1073/PNAS.172226299>.
- (15) Wang, J.; Xie, J.; Schultz, P. G. A Genetically Encoded Fluorescent Amino Acid. *J Am Chem Soc* **2006**, *128* (27), 8738–8739. <https://doi.org/10.1021/ja062666k>.
- (16) Xiu, X.; Puskar, N. L.; Shanata, J. A. P.; Lester, H. A.; Dougherty, D. A. Nicotine Binding to Brain Receptors Requires a Strong Cation–Pi Interaction. *Nature* **2009**, *458*, 534–537. <https://doi.org/10.1038/nature07768>.
- (17) Lieblich, S. A.; Fang, K. Y.; Cahn, J. K. B.; Rawson, J.; LeBon, J.; Teresa Ku, H.; Tirrell, D. A. 4S-Hydroxylation of Insulin at ProB28 Accelerates Hexamer Dissociation and Delays Fibrillation. *J Am Chem Soc* **2017**, *139*, 8384–8387. <https://doi.org/10.1021/jacs.7b00794>.
- (18) Fang, K. Y.; Lieblich, S. A.; Tirrell, D. A. Replacement of ProB28 by Pipecolic Acid Protects Insulin against Fibrillation and Slows Hexamer Dissociation. *J Polym Sci A Polym Chem* **2019**, *57* (3), 264–267. <https://doi.org/10.1002/pola.29225>.
- (19) Deming, T. J.; Fournier, M. J.; Mason, T. L.; Tirrell, D. A. Biosynthetic Incorporation and Chemical Modification of Alkene Functionality in Genetically Engineered Polymers. *J Macromol Sci A* **1997**, *34* (10), 2143–2150. <https://doi.org/10.1080/10601329708010331>.
- (20) Verhoork, S. J. M.; Killoran, P. M.; Coxon, C. R. Fluorinated Prolines as Conformational Tools and Reporters for Peptide and Protein Chemistry. *Biochemistry* **2018**, *57* (43), 6132–6143. <https://doi.org/10.1021/acs.biochem.8b00787>.
- (21) Torbeev, V. Y.; Hilvert, D. Both the *Cis-Trans* Equilibrium and Isomerization Dynamics of a Single Proline Amide Modulate β 2-Microglobulin Amyloid Assembly. *Proc Natl Acad Sci* **2013**, *110* (50), 20051–20056. <https://doi.org/10.1073/pnas.1310414110>.
- (22) Ganguly, H. K.; Basu, G. Conformational Landscape of Substituted Prolines. *Biophys Rev* **2020**, *12*, 25–39. <https://doi.org/10.1007/s12551-020-00621-8>.
- (23) Renner, C.; Alefelder, S.; Bae, J. H.; Budisa, N.; Huber, R.; Moroder, L. Fluoroprolines as Tools for Protein Design and Engineering. *Angew Chem Int Ed* **2001**, *40* (5), 923–925.
- (24) Lieblich, S. Non-Canonical Amino Acid Mutagenesis of Position B28 in Insulin with Proline Analogs, 2017. <https://doi.org/10.7907/Z90Z71BD>.
- (25) Kim, W.; Hardcastle, K. I.; Conticello, V. P. Fluoroproline Flip-Flop: Regiochemical Reversal of a Stereoelectronic Effect on Peptide and Protein Structures. *Angew Chem Int Ed* **2006**, *45* (48), 8141–8145. <https://doi.org/10.1002/anie.200603227>.
- (26) Caporale, A.; Loughlin, J. O.; Ortin, Y.; Rubini, M. A Convenient Synthetic Route to (2*S*,4*S*)-Methylproline and Its Exploration for Protein Engineering of Thioredoxin. *Org Biomol Chem* **2022**. <https://doi.org/10.1039/D2OB01011A>.

- (27) Shoulders, M. D.; Hodges, J. A.; Raines, R. T. Reciprocity of Steric and Stereoelectronic Effects in the Collagen Triple Helix. *J Am Chem Soc* **2006**, *128*, 8112–8113. <https://doi.org/10.1021/ja061793d>.
- (28) Cadamuro, S. A.; Reichold, R.; Kusebauch, U.; Musiol, H. J.; Renner, C.; Tavan, P.; Moroder, L. Conformational Properties of 4-Mercaptoproline and Related Derivatives. *Angew Chem Int Ed* **2008**, *47* (11), 2143–2146. <https://doi.org/10.1002/anie.200704310>.
- (29) Siebler, C.; Erdmann, R. S.; Wennemers, H. Switchable Proline Derivatives: Tuning the Conformational Stability of the Collagen Triple Helix by pH Changes. *Angew Chem Int Ed* **2014**, *53* (39), 10340–10344. <https://doi.org/10.1002/anie.201404935>.
- (30) Torbeev, V. Y.; Fumi, E.; Ebert, M. O.; Schweizer, W. B.; Hilvert, D. Cis-Trans Peptide-Bond Isomerization in α -Methylproline Derivatives. *Helv Chim Acta* **2012**, *95* (12), 2411–2420. <https://doi.org/10.1002/hlca.201200483>.
- (31) De Poli, M.; Moretto, A.; Crisma, M.; Peggion, C.; Formaggio, F.; Kaptein, B.; Broxterman, Q. B.; Toniolo, C. Is the Backbone Conformation of α -Methyl Proline Restricted to a Single Region? *Chem Eur J* **2009**, *15* (32), 8015–8025. <https://doi.org/10.1002/CHEM.200900688>.
- (32) Kim, W.; McMillan, R. A.; Snyder, J. P.; Conticello, V. P. A Stereoelectronic Effect on Turn Formation Due to Proline Substitution in Elastin-Mimetic Polypeptides. *J Am Chem Soc* **2005**, *127* (51), 18121–18132. <https://doi.org/10.1021/ja054105j>.
- (33) Naduthambi, D.; Zondlo, N. J. Stereoelectronic Tuning of the Structure and Stability of the Trp Cage Miniprotein. *J Am Chem Soc* **2006**, *128*, 12430–12431. <https://doi.org/10.1021/ja0648458>.
- (34) Arnold, U.; Raines, R. T. Replacing a Single Atom Accelerates the Folding of a Protein and Increases Its Thermostability. *Org Biomol Chem* **2016**, *14* (28), 6780–6785. <https://doi.org/10.1039/c6ob00980h>.
- (35) Crespo, M. D.; Rubini, M. Rational Design of Protein Stability: Effect of (2S,4R)-4-Fluoroproline on the Stability and Folding Pathway of Ubiquitin. *PLoS One* **2011**, *6* (5), e19425. <https://doi.org/10.1371/journal.pone.0019425>.
- (36) Renner, C.; Alefelder, S.; Bae, J. H.; Budisa, N.; Huber, R.; Moroder, L. Fluoroprolines as Tools for Protein Design and Engineering. *Angew Chem Int Ed* **2001**, *40* (5), 923–925. [https://doi.org/10.1002/1521-3773\(20010302\)40:5<923::AID-ANIE923>3.0.CO;2-#](https://doi.org/10.1002/1521-3773(20010302)40:5<923::AID-ANIE923>3.0.CO;2-#).
- (37) Zheng, T. Y.; Lin, Y. J.; Horng, J. C. Thermodynamic Consequences of Incorporating 4-Substituted Proline Derivatives into a Small Helical Protein. *Biochemistry* **2010**, *49* (19), 4255–4263. <https://doi.org/10.1021/bi100323v>.
- (38) Dietz, D.; Kubyshkin, V.; Budisa, N. Applying γ -Substituted Prolines in the Foldon Peptide: Polarity Contradicts Preorganization. *ChemBioChem* **2015**, *16* (3), 403–406. <https://doi.org/10.1002/cbic.201402654>.
- (39) Rubini, M.; Schärer, M. A.; Capitani, G.; Glockshuber, R. (4R)- and (4S)-Fluoroproline in the Conserved *Cis*-Prolyl Peptide Bond of the Thioredoxin Fold: Tertiary Structure

- Context Dictates Ring Puckering. *ChemBioChem* **2013**, *14* (9), 1053–1057. <https://doi.org/10.1002/cbic.201300178>.
- (40) Roderer, D.; Lockshuber, R.; Rubini, M. Acceleration of the Rate-Limiting Step of Thioredoxin Folding by Replacement of Its Conserved *Cis*-Proline with (4*S*)-Fluoroproline. *ChemBioChem* **2015**, *16*, 2162–2166. <https://doi.org/10.1002/cbic.201500342>.
- (41) O' Loughlin, J.; Napolitano, S.; Rubini, M. Protein Design with Fluoroprolines: 4,4-Difluoroproline Does Not Eliminate the Rate-Limiting Step of Thioredoxin Folding. *ChemBioChem* **2021**, *22* (23), 3326–3332. <https://doi.org/10.1002/CBIC.202100418>.
- (42) Torbeev, V.; Ebert, M. O.; Dolenc, J.; Hilvert, D. Substitution of Proline32 by α -Methylproline Preorganizes β 2-Microglobulin for Oligomerization but Not for Aggregation into Amyloids. *J Am Chem Soc* **2015**, *137* (7), 2524–2535. <https://doi.org/10.1021/ja510109p>.
- (43) Steiner, T.; Hess, P.; Bae, J. H.; Wiltschi, B.; Moroder, L.; Budisa, N. Synthetic Biology of Proteins: Tuning GFPs Folding and Stability with Fluoroproline. *PLoS One* **2008**, *3* (2), e1680. <https://doi.org/10.1371/journal.pone.0001680>.
- (44) Deepankumar, K.; Nadarajan, S. P.; Ayyadurai, N.; Yun, H. Enhancing the Biophysical Properties of MRFP1 through Incorporation of Fluoroproline. *Biochem Biophys Res Commun* **2013**, *440* (4), 509–514. <https://doi.org/10.1016/j.bbrc.2013.09.062>.
- (45) Edwardraja, S.; Sriram, S.; Govindan, R.; Budisa, N.; Lee, S.-G. Enhancing the Thermal Stability of a Single-Chain Fv Fragment by in Vivo Global Fluorination of the Proline Residues. *Mol BioSyst* **2011**, *7*, 258–265. <https://doi.org/10.1039/c0mb00154f>.
- (46) Merkel, L.; Schauer, M.; Antranikian, G.; Budisa, N. Parallel Incorporation of Different Fluorinated Amino Acids: On the Way to “Teflon” Proteins. *ChemBioChem* **2010**, *11* (11), 1505–1507. <https://doi.org/10.1002/cbic.201000295>.
- (47) Holzberger, B.; Marx, A. Replacing 32 Proline Residues by a Noncanonical Amino Acid Results in a Highly Active DNA Polymerase. *J Am Chem Soc* **2010**, *132* (44), 15708–15713. <https://doi.org/10.1021/ja106525y>.
- (48) Holzberger, B.; Obeid, S.; Welte, W.; Diederichs, K.; Marx, A. Structural Insights into the Potential of 4-Fluoroproline to Modulate Biophysical Properties of Proteins. *Chem Sci* **2012**, *3* (10), 2924–2931. <https://doi.org/10.1039/c2sc20545a>.
- (49) Lajoie, M. J.; Rovner, A. J.; Goodman, D. B.; Aerni, H.; Haimovich, A. D.; Kuznetsov, G.; Mercer, J. A.; Wang, H. H.; Carr, P. A.; Mosberg, J. A.; Rohland, N.; Schultz, P. G.; Jacobson, J. M.; Rinehart, J.; Church, G. M.; Isaacs, F. J. Genomically Recoded Organisms Expand Biological Functions. *Science (1979)* **2013**, *342* (October), 357–360. <https://doi.org/10.1126/science.1241459>.
- (50) Dunkelmann, D. L.; Willis, J. C. W.; Beattie, A. T.; Chin, J. W. Engineered Triply Orthogonal Pyrrolysyl-tRNA Synthetase/tRNA Pairs Enable the Genetic Encoding of Three Distinct Non-Canonical Amino Acids. *Nat Chem* **2020**, *12* (6), 535–544. <https://doi.org/10.1038/s41557-020-0472-x>.

- (51) Sakamoto, K.; Hayashi, A.; Sakamoto, A.; Kiga, D.; Nakayama, H.; Soma, A.; Kobayashi, T.; Kitabatake, M.; Takio, K.; Saito, K.; Shirouzu, M.; Hirao, I.; Yokoyama, S. Site-Specific Incorporation of an Unnatural Amino Acid into Proteins in Mammalian Cells. *Nucleic Acids Res* **2002**, *30* (21), 4692–4699. <https://doi.org/10.1093/NAR/GKF589>.
- (52) Furter, R. Expansion of the Genetic Code: Site-Directed p-Fluoro-Phenylalanine Incorporation in *Escherichia coli*. *Protein Sci* **1998**, 7419–7426. <https://doi.org/10.1002/pro.5560070223>.
- (53) Kiick, K. L.; van Hest, J. C. M.; Tirrell, D. A. Expanding the Scope of Protein Biosynthesis by Altering the Methionyl-tRNA Synthetase Activity of a Bacterial Expression Host. *Angew Chem Int Ed* **2000**, *39* (12), 2148–2152. [https://doi.org/10.1002/1521-3757\(20000616\)112:12<2232::AID-ANGE2232>3.0.CO;2-0](https://doi.org/10.1002/1521-3757(20000616)112:12<2232::AID-ANGE2232>3.0.CO;2-0).
- (54) Tanrikulu, I. C.; Schmitt, E.; Mechulam, Y.; Goddard, W. A.; Tirrell, D. A. Discovery of *Escherichia coli* Methionyl-tRNA Synthetase Mutants for Efficient Labeling of Proteins with Azidonorleucine in Vivo. *Proc Natl Acad Sci* **2009**, *106* (36), 15285–15290. <https://doi.org/10.1073/pnas.0905735106>.
- (55) Teeuwen, R. L. M.; Van Berkel, S. S.; Van Dulmen, T. H. H.; Schoffelen, S.; Meeuwissen, S. A.; Zuilhof, H.; De Wolf, F. A.; Van Hest, J. C. M. “Clickable” Elastins: Elastin-like Polypeptides Functionalized with Azide or Alkyne Groups. *Chem Commun* **2009**, No. 27, 4022–4024. <https://doi.org/10.1039/B903903A>.
- (56) Wolschner, C.; Giese, A.; Kretzschmar, H. A.; Huber, R.; Moroder, L.; Budisa, N. Design of Anti- and pro-Aggregation Variants to Assess the Effects of Methionine Oxidation in Human Prion Protein. *Proc Natl Acad Sci* **2009**, *106* (19), 7756–7761. <https://doi.org/10.1073/PNAS.0902688106>.
- (57) Montclare, J. K.; Son, S.; Clark, G. A.; Kumar, K.; Tirrell, D. A. Biosynthesis and Stability of Coiled-Coil Peptides Containing (2*S*,4*R*)-5,5,5-Trifluoroleucine and (2*S*,4*S*)-5,5,5-Trifluoroleucine. *ChemBioChem* **2009**, *10* (1), 84–86. <https://doi.org/10.1002/CBIC.200800164>.
- (58) Yoo, T. H.; Link, A. J.; Tirrell, D. A. Evolution of a Fluorinated Green Fluorescent Protein. *Proc Natl Acad Sci* **2007**, *104* (35), 13887–13890. <https://doi.org/10.1073/pnas.0701904104>.
- (59) Budisa, N.; Minks, C.; Medrano, F. J.; Lutz, J.; Huber, R.; Moroder, L. Residue-Specific Bioincorporation of Non-Natural, Biologically Active Amino Acids into Proteins as Possible Drug Carriers: Structure and Stability of the per-Thiaproline Mutant of Annexin V. *Proc Natl Acad Sci* **1998**, *95* (2), 455–459. <https://doi.org/10.1073/pnas.95.2.455>.
- (60) Kim, W.; George, A.; Evans, M.; Conticello, V. P. Cotranslational Incorporation of a Structurally Diverse Series of Proline Analogues in an *Escherichia coli* Expression System. *ChemBioChem* **2004**, *5* (7), 928–936. <https://doi.org/10.1002/cbic.200400052>.

- (61) Lukesch, M. S.; Pavkov-Keller, T.; Gruber, K.; Zangger, K.; Wiltschi, B. Substituting the Catalytic Proline of 4-Oxalocrotonate Tautomerase with Non-Canonical Analogues Reveals a Finely Tuned Catalytic System. *Sci Rep* **2019**, *9*, 2697. <https://doi.org/10.1038/s41598-019-39484-9>.
- (62) Grothe, S.; Krogsrud, R. L.; McClellan, D. J.; Milner, J. L.; Wood, J. M. Proline Transport and Osmotic Stress Response in *Escherichia coli* K-12. *J Bacteriol* **1986**, *166* (1), 253–259.
- (63) Fang, K. Y. Modulating Biophysical Properties of Insulin with Non-Canonical Mutagenesis at Position B28, California Institute of Technology, 2017. <https://doi.org/10.7907/Z9V40S6J>.

*Chapter II*INCORPORATION OF ALIPHATIC PROLINE ANALOGS INTO
RECOMBINANTLY-PRODUCED INSULIN**2.1 Contributions**

Chad D. Paavola, Michael P. Akers, Julie S. Moyers, and Alborz Mahdavi provided advice on proinsulin refolding. Alejandro Lopez assisted with cloning *proS* mutants, and Alex M. Chapman performed pilot ncPro incorporation studies. Janine Quijano and Cecile Donohue performed mouse experiments, with input from Hsun Teresa Ku. Analytical ultracentrifugation was performed by Amy Henrickson and Maduni Ranasinghe, with input from Borries Demeler. Jay Winkler, Songye Chen, and Mona Shahgholi assisted with circular dichroism spectroscopy, transmission electron microscopy, and mass spectrometry, respectively. Stephanie L. Breunig prepared all insulin samples and performed all other experiments.

2.2 Abstract

Non-canonical proline residues expand the chemical space about proline while maintaining its conformationally-restricting properties. Here, we demonstrate the incorporation of 4*R*-methylproline, 4*S*-methylproline, and 4-methyleneproline into recombinantly-produced insulin in *Escherichia coli* using a residue-specific replacement approach. These modest modifications to proline at position B28 affect the biophysical properties of insulin: incorporation of 4-methyleneproline at B28 accelerates fibril formation, while 4-methylation speeds dissociation from the pharmaceutically-formulated hexamer. This

work expands the scope of proline analogs amenable to incorporation into recombinant proteins, and demonstrates how non-canonical amino acid mutagenesis can engineer the therapeutically-relevant properties of protein drugs.

2.3 Introduction

Proline is unique among the canonical amino acids: the cyclic pyrrolidine side chain restricts the conformational space accessible to the residue. Replacing proline with any of the proteinogenic amino acids through standard mutagenesis approaches necessarily grants greater conformational freedom. Alternatively, non-canonical proline (ncPro) residues expand the chemical space about proline, while maintaining a pyrrolidine (or pyrrolidine-like) side chain. Because the conformational preferences of many proline analogs are known,^{1,2} ncPro residues can also probe the effect that a particular proline conformation (i.e., *cis* or *trans* amide isomer) has on protein behavior. Proline analogs have demonstrated the importance of a key proline *cis-trans* isomerization event in 5HT₃ receptor opening,³ modified the properties of elastin-like proteins,⁴⁻⁶ determined the molecular origins of collagen stability,^{1,7} and probed the role of *cis-trans* isomerization in β 2-microglobulin fibrillation.^{8,9} To date, residue-specific (rather than site-specific) incorporation approaches remain the only method for installing proline analogs into recombinant proteins. Using *E. coli* as an expression host, this approach leads to the efficient incorporation of diverse 3- and 4-functionalized proline residues, and those with modified ring sizes and compositions.¹⁰

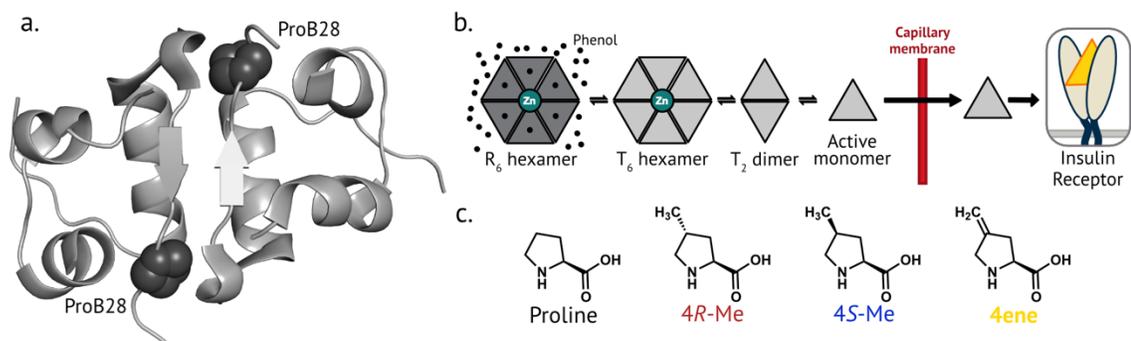


Figure 2.1. Proline mutagenesis at position B28 of human insulin. **a.** Crystal structure of insulin (PDB 1MSO), highlighting ProB28 located at the dimer interface. **b.** Simplified scheme of insulin dissociation after injection. Insulin exists as a hexamer in the R state in the presence of zinc and phenolic ligands, such as in the pharmaceutical formulation. After injection, insulin dissociates into lower-order oligomeric species that can more easily diffuse across the capillary membrane, enter the bloodstream, and bind to the insulin receptor. **c.** The structure of proline and the aliphatic proline analogs used in this study.

Insulin (Figure 2.1a) is a 5.8 kDa peptide hormone normally released from the pancreatic β cells in response to elevated levels of blood glucose. Its binding to the insulin receptor induces several intracellular responses that ultimately lower blood glucose concentrations.¹¹ Diabetes mellitus is a result of dysfunctional insulin signaling, either through an impaired ability to produce and secrete insulin (type 1), or by insulin resistance (type 2). Subcutaneous injection of exogenous insulin is a common strategy in diabetes treatment, especially for individuals with type 1 diabetes. The recombinant production of insulin¹² has significantly advanced diabetes treatment by enabling both its production at scale, and the creation of modified insulins with desirable properties through standard mutagenesis approaches.^{13–16}

To more closely mimic the insulin-action profile of a healthy pancreas, two broad classes of insulin analogs have been developed: long-acting (or basal) insulins, and fast-acting insulins (FAIs).¹⁷ Long-acting variants recapitulate the lower levels of insulin secretion that maintain metabolism in an anabolic state. Conversely, FAIs aim to mimic the transient increases in insulin concentration stimulated by elevated blood glucose, such as after a meal. Typical insulin replacement therapy relies on a combination of regular basal insulin treatments, and FAI injections before meals.

Many FDA-approved insulin variants contain minimal alterations to insulin's amino acid sequence that lead to pronounced pharmacokinetic effects.¹⁸ Notably, insulin aspart^{14,19} (NovoLog, marketed by Novo Nordisk) and insulin lispro^{15,20} (Humalog, Eli Lilly) both involve changes to ProB28, a key dimer-stabilizing residue¹⁵ near the C-terminus of the B-chain (Figure 2.1a). Insulin aspart is achieved by the single point mutation ProB28Asp, while insulin lispro contains an inversion of ProB28 and LysB29; both changes destabilize oligomer formation. Because the rate-limiting step for insulin absorption into the bloodstream is dissociation of oligomer to monomer²¹ (Figure 2.1b), these changes accelerate insulin's onset of action. Insulins are also prone to chemical and physical denaturation,²²⁻²⁴ processes slowed by the formation of protective oligomers. As a result, insulin production and distribution require a cold chain,²⁵ and its long-term storage in continuous subcutaneous insulin infusion pumps is challenging.²⁶

Intrigued by the role that ProB28 plays in insulin biophysics, we recently introduced ncPro residues at position B28 of human insulin.^{27,28} Because mature insulin contains one proline,

a residue-specific replacement approach results in site-specific proline replacement without the need for an orthogonal aminoacyl-tRNA synthetase/tRNA pair. These efforts, which focused on proline analogs known to incorporate well into recombinant proteins in *E. coli*,⁵ illustrated how proline mutagenesis of insulin can tune its biophysical characteristics.^{27,28}

Here, we demonstrate the efficient incorporation of three new aliphatic proline residues (4*R*-methylproline, 4*R*-Me; 4*S*-methylproline, 4*S*-Me; and 4-methyleneproline, 4ene; Figure 2.1c) at position B28 of recombinantly-produced insulin; these insulin variants will be referred to as ins-4*R*-Me, ins-4*S*-Me, and ins-4ene, respectively. We find that these modifications alter insulin behavior: replacement of ProB28 with 4ene speeds fibril formation, while 4-methylation accelerates hexamer dissociation. This work expands the range of proline analogs that can be incorporated into recombinant proteins in *E. coli*. It also demonstrates how small molecular changes introduced through non-canonical amino acid mutagenesis can be used to probe and engineer the therapeutically-relevant properties of protein drugs.

2.4 Results and discussion

2.4.1 Aliphatic proline residues are accepted by the E. coli translational machinery

To identify an expanded set of ncPro residues accepted by the *E. coli* translational machinery, we expressed proinsulin (a precursor to insulin) under conditions that favor ncPro incorporation.¹⁰ We monitored ncPro replacement by proinsulin expression and mass spectrometry, and noted a range of incorporation efficiencies for the 15

commercially-available proline analogs tested (Figure 2.S1, Table 2.S1). Notably, the aliphatic proline residues 4*R*-Me, 4*S*-Me, and 4ene led to high levels of proinsulin expression and good (~90%) incorporation efficiencies under optimized conditions (Figure 2.2a-d, Table 2.S2, Table 2.S3). We chose to replace ProB28 of mature human insulin with each of these three aliphatic proline analogs (Figure 2.2e-h), and determined the effect that aliphatic proline replacement has on insulin behavior.

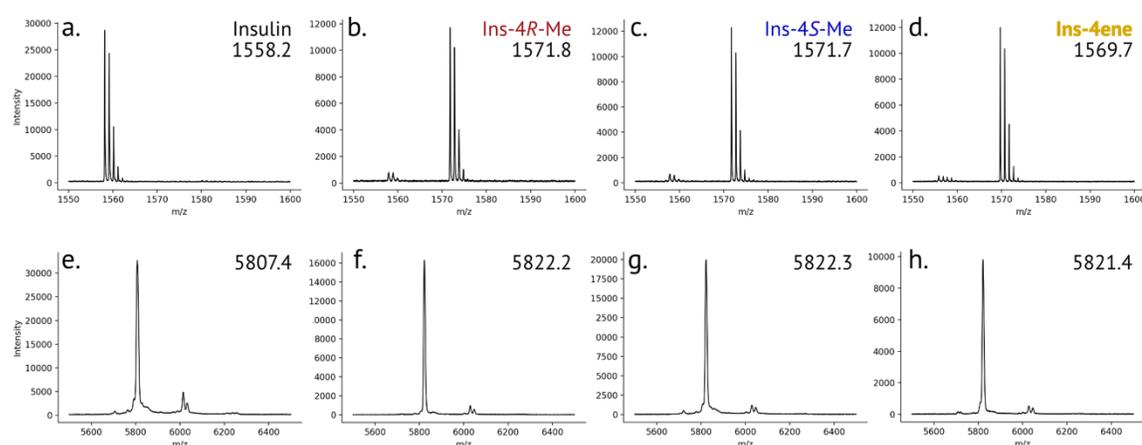


Figure 2.2. Mass spectrometry characterization of insulin variants. a-d. Characterization of proline analog incorporation. The solubilized inclusion body (containing proinsulin) after expression in the presence of proline (a), 4*R*-Me (b), 4*S*-Me (c), 4ene (d) was digested with Glu-C and analyzed by MALDI-TOF MS. The peptide that contains position B28 of mature insulin is ⁵⁰RGFFYTPKTRRE (expected $m/z = 1557.8$). **e-h.** MALDI-TOF characterization of mature and purified insulin variants: human insulin (e), Ins-4*R*-Me (f), Ins-4*S*-Me (g), and Ins-4ene (h). The larger molecular weight peaks ($m/z \sim 6050$) present in these spectra correspond to adducts of the sinapic acid matrix.

2.4.2 Proline analogs do not affect insulin secondary structure or bioactivity

The secondary structure of each insulin variant was assessed by circular dichroism spectroscopy (Figure 2.3a-c; Table 2.S4). CD can assess insulin oligomerization: for

monomeric insulins, the ratio of negative ellipticities at 208 and 222 nm is increased compared to that of the insulin dimer.²⁹ The far-UV CD spectrum for each variant closely matched that of human insulin at 60 μM , suggesting that proline replacement has not significantly perturbed secondary structure or dimer formation under these conditions.

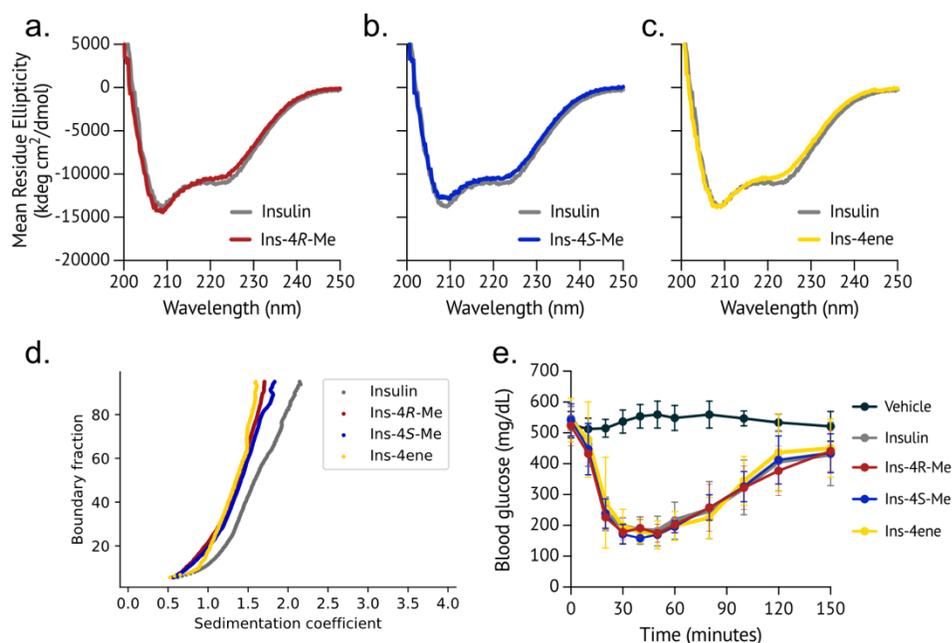


Figure 2.3. Circular dichroism spectroscopy, sedimentation, and bioactivity of insulin variants. **a-c.** Far-UV circular dichroism spectra of insulin and insulin variants (60 μM in 100 mM phosphate buffer, pH 8.0). The spectrum of each insulin variant is overlaid with that of human insulin (grey). **d.** Velocity sedimentation of insulin and insulin variants (60 μM in 100 mM phosphate buffer, pH 8.0). **e.** Insulins were injected subcutaneously into diabetic mice, and blood glucose was measured over time after injection.

Insulin oligomerization was studied further by velocity sedimentation analytical ultracentrifugation (Figure 2.3d, Table 2.S4). The broad sedimentation profiles of each insulin under these conditions (60 μM insulin, 100 mM phosphate, pH 8.0) are suggestive of a mixture of oligomeric states (here, predominantly monomers and dimers). The

sedimentation coefficients of all insulin variants (1.3-1.4S) were modestly reduced compared to that of human insulin (1.6S). These smaller sedimentation coefficients might be attributed to increased monomer-dimer dissociation constants. The K_d of Ins-4ene was determined from these data (25 μ M, 95% CI [22, 28]), and is indeed moderately increased compared to the reported monomer-dimer K_d of human insulin (9 μ M).³⁰ However, it was difficult to accurately determine the dissociation constants for the other insulins due to the existence of higher-order oligomers in the samples.

To validate biological activity *in vivo*, insulins were formulated with zinc and phenolic ligands, and injected subcutaneously into diabetic mice; blood glucose was monitored over 2.5 h. Rodent models can assess insulin activity, but cannot distinguish differences in onset of action between human insulin and fast-acting analogs.³¹ Since the B-chain C-terminus does not interact with the insulin receptor,^{32,33} we did not expect modification of ProB28 to hinder insulin's bioactivity. Indeed, all insulin variants reduced blood glucose in diabetic mice (Figure 2.3e).

2.4.3 4-methylene at position B28 accelerates fibril formation

To assess stability against physical denaturation, insulins were subjected to vigorous shaking at 37°C, and fibril formation was monitored over time with the fibril-specific dye thioflavin T (ThT; Figure 2.4a; Table 2.S4). The introduction of either 4*R*-Me or 4*S*-Me at insulin B28 did not significantly affect the lag time to fibril formation under these experimental conditions, compared to human insulin (lag time = 16.6 \pm 4.1 h). However, ins-4ene (8.2 \pm 4.0 h) formed fibrils more rapidly than the other insulins assessed here. All

aggregates were fibrillar in nature when assessed by transmission electron microscopy (TEM). Compared to the micrometer-long fibrils described in previous reports,^{34–37} those observed here were relatively short (tens to hundreds of nanometers). Under these conditions, ins-4ene formed shorter fibrils compared to the other insulins (Figure 2.4b-e).

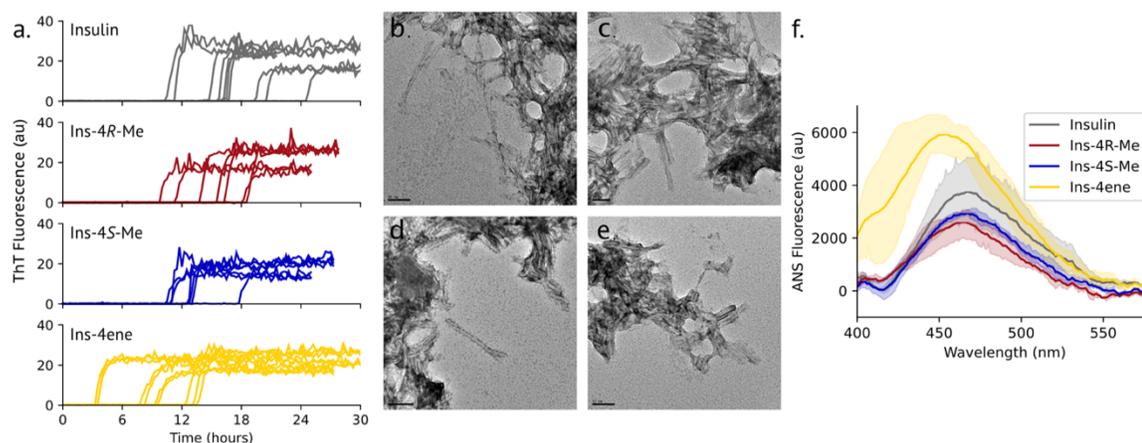


Figure 2.4. Fibrillation of insulin variants **a.** Insulin variants (60 μM in 100 mM phosphate buffer, pH 8.0) were incubated at 37°C with vigorous shaking, and fibril formation was monitored by ThT fluorescence over time. **b-e.** Representative TEM images of insulin (**b**; scale bar: 50 μm), ins-4R-Me (**c**; 50 μm), ins-4S-Me (**d**; 100 μm), and ins-4ene (**e**; 50 μm) aggregates. **f.** ANS emission spectra of insulin variants (1 μM insulin variant labeled with 5 μM ANS in 100 mM phosphate buffer, pH 8.0).

We sought to better understand the molecular mechanism for the decreased stability of ins-4ene. An early step in the mechanism of insulin fibril formation is thought to be detachment of the C-terminus of the insulin B-chain from the core of the molecule, exposing hydrophobic residues.³⁷ We probed insulin disorder with the dye 8-anilino-1-naphthalenesulfonic acid (ANS).³⁸ Compared to the other insulins measured, ins-4ene exhibits a blue-shift in the emission maximum and increase in fluorescence intensity upon labeling with ANS (Figure 2.4f; Table 2.S4), suggesting greater disorder. We analyzed

DMSO-solubilized ins-4ene fibrils by mass spectrometry, but did not observe chemical modification of the exocyclic alkene (data not shown).

2.4.4 4-methylation of ProB28 speeds hexamer dissociation

Increased negative ellipticity at 222 nm is indicative of insulin oligomerization.^{29,39} Inspired by a previous report,⁴⁰ we measured the rate of dissociation to the monomer state using CD spectroscopy. Insulins were formulated under conditions that mimic the pharmaceutical formulation (600 μ M insulin, 25 mM *m*-cresol, 250 μ M ZnCl₂) and favor the R₆ hexamer state.⁴¹ Dissociation was monitored by tracking mean residue ellipticity at 222 nm over time after 150-fold dilution into ligand-free buffer (Figure 2.5a). CD spectra after dilution are distinct from that of chemically-denatured insulin (Figure 2.S3), suggesting that changes in CD signal are not a result of denaturation. While ins-4ene dissociated similarly to human insulin ($t_{1/2} = 17.0 \pm 2.3$ s and 18.4 ± 2.8 s, respectively), the dissociation kinetics of ins-4*R*-Me ($t_{1/2} = 9.8 \pm 1.8$ s) and ins-4*S*-Me ($t_{1/2} = 9.9 \pm 2.5$ s) were accelerated under these conditions (Figure 2.5b-f, Table 2.S4).

Steric effects might play a role in accelerating the dissociation of ins-4*R*-Me and ins-4*S*-Me. A 4*R*- or 4*S*-methyl substituent installed at ProB28 in published structures of the R₆ insulin hexamer are in close proximity (2.2 and 2.3 Å, respectively) to backbone carbonyl oxygen atoms of the adjacent monomer (Figure 2.S4a); these distances are within the van der Waals radii of the respective atoms. Similar interactions are observed in the T₆ hexamer (Figure 2.S4b). The resulting steric clashes perhaps destabilize the hexamer and accelerate dissociation.

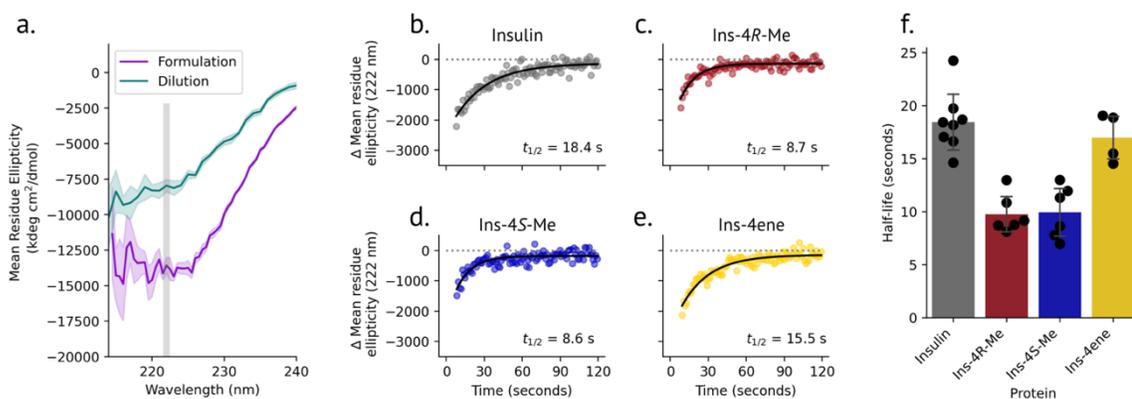


Figure 2.5. Hexamer dissociation kinetics of insulin variants **a.** Equilibrium CD spectra of insulin before and after dilution. To measure dissociation kinetics, the decrease in negative ellipticity at 222 nm was monitored over time after dilution. **b-e.** Representative dissociation kinetics for insulin (**b**), ins-4R-Me (**c**), ins-4S-Me (**d**), and ins-4ene (**e**). Each dilution experiment was fit to a mono-exponential function, and the half-life for each displayed replicate is indicated. **f.** Summary of dissociation half-life values.

We examined the oligomerization of insulin and ins-4S-Me in more detail by sedimentation velocity analytical ultracentrifugation. At 300 μ M and in the presence of zinc and *m*-cresol, the sedimentation coefficients of both insulin (3.3S) and ins-4SMe (3.4S) correspond to the hexamer⁴² (Figure 2.6a, Table 2.S5). After dilution, both insulins behave as monomers (Figure 2.6b, Table 2.S5), further validating that the dilution kinetics measured above indeed describe hexamer dissociation. Interestingly, we found that 4S-methylation of ProB28 affects insulin oligomerization in the absence of ligands. Compared to the broad sedimentation profile of insulin, which is indicative of insulin's heterogeneous oligomerization behavior,⁴³ the sedimentation of ins-4S-Me (3.1S) is nearly unchanged in the absence of zinc and *m*-cresol (Figure 2.6b).

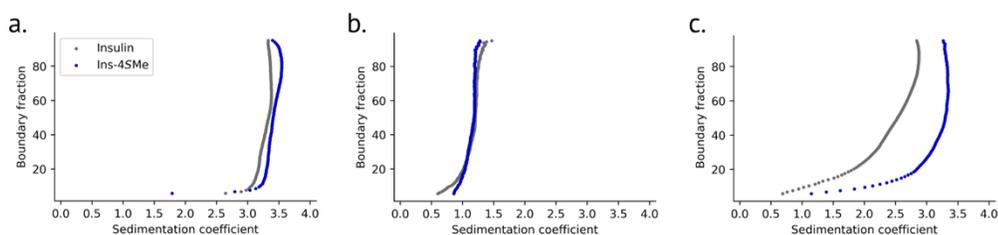


Figure 2.6. Sedimentation velocity of insulin and ins-4S-Me. Insulins were formulated under the following conditions: **a.** 300 μM insulin, 12.5 mM m-cresol, 125 μM Zn. **b.** 4 μM insulin, 167 μM m-cresol, 1.67 μM Zn. **c.** 300 μM insulin. All samples were in 25 mM tris buffer, pH 8.0.

2.5 Conclusion

Here, we have expanded the set of ncPro residues that can be incorporated into proteins in *E. coli*, and demonstrated their utility in modifying the properties of recombinantly-produced insulin. We found that the proline analogs 4*R*-Me, 4*S*-Me, and 4ene could be incorporated with high efficiencies into recombinant proinsulin. During the completion of this work, the incorporation of 4*R*-Me was also reported in the recombinant expression of thioredoxin.⁴⁴ In that case, incorporation of the 4*S* diastereomer could not be detected, perhaps due to its interference with thioredoxin folding. Conversely, insoluble proinsulin is purified from the inclusion body fraction, so any influence of ncPro replacement on protein stability during expression is likely reduced. We were also able to detect low to modest levels of incorporation of other ncPro residues, such as the diazirine-containing variant photo-proline, and 3-hydroxy and 4-oxo analogs (Figure 2.S1, Table 2.S1). Engineering the prolyl-tRNA synthetase or other components of the *E. coli* translational machinery might lead to improved incorporation of these analogs.

Besides expanding the size and hydrophobic surface area at proline residues, the three aliphatic ncPro residues studied here complement other translationally-active proline

analogues used to probe proline conformation. For instance, 4*R*-Me and 4*S*-Me have opposing conformational preferences relative to their more commonly-used 4-fluoroproline counterparts,^{1,2,45} and together have been used to validate the stereoelectronic origin of collagen stability.^{7,45} The amplitude of the pyrrolidine ring pucker of 4ene is expected to be attenuated compared to proline (Figure 2.S5), a conformation approaching that of the planar 3,4-dehydroproline.⁴⁶ Furthermore, the olefin present in 4ene might be used as a chemical handle for protein modification at proline residues.^{47,48}

This work also demonstrates how small molecular changes can have oversized effects on protein behavior. For instance, introducing an exocyclic alkene at B28 of insulin shortened its lag time to fibril formation. This finding, together with increased disorder of ins-4ene as evidenced by its ANS emission spectrum, supports current models for the mechanism of insulin fibrillation.³⁷ Furthermore, installing a simple methyl group at ProB28 of human insulin significantly accelerates dissociation from the insulin hexamer, behavior associated with achieving a rapid onset-of-action for FAIs.⁴⁹ Future work is needed to more fully understand the effect that 4-methylation has on insulin oligomerization, and to determine if the faster dissociation kinetics observed for ins-4*R*-Me and ins-4*S*-Me *in vitro* result in a faster onset-of-action *in vivo*.

2.6 Materials and methods

2.6.1 Chemicals

All chemicals were purchased from MilliporeSigma unless otherwise indicated. 4-methyleneproline (4ene) was purchased as the N-boc protected version from Acros

Organics, and deprotected with trifluoroacetic acid (TFA) in dichloromethane. 4ene was extracted with H₂O and lyophilized; complete deprotection and >95% purity was verified by ¹H NMR. All other proline analogs were used as received: 4*R*-methylproline (4*R*-Me) was purchased from Advanced Chemblocks as the hydrochloride salt; ¹H NMR analysis indicated the presence of 5-10% of the 4*S*-methylproline diastereomer. 2-methylproline (2-Me) was purchased from Advanced ChemBlocks as the hydrochloride salt. 4*S*-methylproline (4*S*-Me) was purchased from AstaTech as the hydrochloride salt. L-*cis*-pyrrolidine-2,4-dicarboxylic acid (4*S*-COOH) was purchased from Boc Sciences. L-*trans*-pyrrolidine-2,4-dicarboxylic acid (4*R*-COOH) was purchased from Tocris Biosciences. L-pyrroglutamic acid (5-oxo) was purchased from Aldrich. 4,4-dimethylproline (44-diMe) was purchased from J&W Pharmed. Piperazine-2-carboxylic acid (Pip-Az) was purchased from Ark Pharm as the dihydrochloride salt. (2*S*,5*S*)-5-hydroxypiperidine-2-carboxylic acid (Pip-OH) was purchased from Ark Pharm. 4*S*-aminoproline (4*S*-NH₂) was purchased from Toronto Research Chemicals as the dihydrochloride salt. 3*R*-hydroxyproline (3*R*-OH) was purchased from Combi-Blocks. 3*S*-hydroxyproline (3*S*-OH) was purchased from Ark Pharm. 4-oxoproline (4-oxo) was purchased as the hydrobromide salt from MilliporeSigma.

2.6.2 Enzymes

Restriction enzymes, kinases, and ligases were purchased from New England Biolabs. Trypsin was purchased from MilliporeSigma. Carboxypeptidase B was purchased from Worthington Biochemical. Glu-C peptidase was purchased from Promega.

2.6.3 Strains and plasmids

The proline-auxotrophic *E. coli* strain CAG18515 was obtained from the Coli Genetic Stock Center (CGSC) at Yale University. Strain DH10B was used for standard cloning operations; electrocompetent CAG18515 were transformed with purified plasmid products.

The plasmid pQE80_H27R-PI_proS contains an IPTG-inducible proinsulin gene and the *E. coli* prolyl-tRNA synthetase gene controlled by its endogenous promoter. Proinsulin is translationally fused to an N-terminal leader peptide (H27R) that increases expression yields,⁵⁰ and a 10x-his tag to facilitate proinsulin enrichment after refolding. The gene for H27R-PI was ordered as a g-Block gene fragment from Integrated DNA Technologies (IDT) after codon optimization of the N-terminal leader peptide. A restriction enzyme cloning approach (XhoI and BamHI restriction enzyme cut sites) was used to replace the hexahistidine-tagged proinsulin gene in the plasmid pQE80PI-proS, which was described previously.²⁸ Correct installation of the gene of interest was verified by Sanger sequencing.

A blunt-end ligation approach was used to install the C443G and M157Q mutations. The *proS*-containing plasmid was amplified with primers AL01004_fwd & AL01004_rev (C443G), or AL01005_fwd & AL01005_rev2 (M157Q). The linear PCR product was phosphorylated (T4 PNK) and circularized (T4 DNA ligase). Correct installation of the point mutation was verified by Sanger sequencing.

2.6.4 Primers

DNA oligos were purchased from Integrated DNA Technologies (IDT). Nucleotides responsible for installing the M157Q and C443G mutations are underlined.

AL01004_fwd: ATACCGTAGCCACCCATCGTCAGG

AL01004_rev: CGGGGTAACGCGTGTGGT

AL01005_fwd: AGCATCTTTCTGCCAGGAATTCGC

AL01005_rev2: TACTCTTTCCATACTTCTCAGGAATCC

2.6.5 Nucleotide and amino acid sequences

H27R-PI: The g-Block gene fragment was purchased from IDT. The coding sequence is in UPPERCASE, XhoI and BamHI cut sites are underlined.

gcccttcgtcttcacctcgagaaatcataaaaaattatttgctttgtgagcggataacaattataatagattcaattgtgagcggata
acaatttcacacagaattcattaaagaggagaaattaactATGACAATGATCACTAATTCACCCGAGA
TTCCACCATCATCATCATCACCACCACCATCAGTTGATCTCGGAGGCC
CGTTTTGTGAACCAGCACCTGTGCGGTAGCCACCTGGTGGAAAGCTCTGTACCT
GGTTTGCGGTGAGCGTGGTTTCTTCTACACGCCAAAGACCCGCCGTGAAGCT
GAAGATCTGCAGGTGGGCCAGGTAGAACTGGGCGGTGGTCCGGGTGCCGGCT
CTCTGCAACCGCTGGCACTGGAAGGTTCCCTGCAAGCGCGTGGTATCGTAGA
GCAGTGCTGTACTTCTATCTGCTCCCTGTACCAGCTGGAGAACTACTGTAATT
AAggatccgcatgcgagc

The sequence of the H27R leader peptide is underlined, proinsulin is in **bold**. The A-chain and B-chain in mature insulin are colored **red** and **blue**, respectively.

MTMITNSPEISHHHHHHHHHHHQLISEAR**FVNQHL****CGSHLVEALYLVCGERGFF**
YTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQARGIVEQCCTSICS
LYQLENYCN

proS: the endogenous *proS* promoter is underlined, the coding sequence is in UPPERCASE.

attcacgccctctcttttgacatttctttgcactggtaaactaaatcactttttttgtcccaggctgccttgagcctgttctaccttc
 aactggaaccgtaacaacATGCGTACTAGCCAATACCTGCTCTCCACTCTCAAGGAGAC
 ACCTGCCGACGCCGAGGTGATCAGCCATCAGCTGATGCTGCGCGCCGGGATG
 ATCCGCAAGCTGGCCTCCGGGTTATATACCTGGCTGCCGACCGGCGTGCGCG
 TTCTGAAAAAAGTCGAAAACATCGTGCGTGAAGAGATGAACAACGCCGGTGC
 GATCGAGGTGTCGATGCCGGTGGTTCAGCCAGCCGATTTGTGGCAAGAGAGT
 GGTCGTTGGGAACAGTACGGTCCGGAAGTCTGCGTTTTGTTGACCGTGGCG
 AGCGTCCGTTTCGTAATCGGCCCAACTCATGAAGAAGTTATCACTGACCTGATT
 CGTAACGAGCTTAGCTCTTACAAACAGCTGCCGCTGAACTTCTATCAGATCCA
 GACCAAGTTCCGCGACGAAGTGCCTCCGCGTTTCGGCGTCATGCGTTCCCGC
 GAATTCCTGATGAAAGATGCTTACTCTTCCATACTTCTCAGGAATCCCTGCA
 GGAAACCTACGATGCAATGTATGCGGCCTACAGCAAAATCTTCAGCCGCATG
 GGGCTGGATTTCCGCGCCGTACAAGCCGACACCGGTTCTATCGGCGGCAGCG
 CCTCTCACGAATTCCAGGTGCTGGCGCAGAGCGGTGAAGACGATGTGGTCTT
 CTCCGACACCTCTGACTATGCAGCGAACATTGAACTGGCAGAAGCTATCGCG
 CCGAAAGAACCGCGCGCTGCTGCTACCCAGGAAATGACGCTGGTTGATACGC
 CGAACGCGAAAACCATCGCGGAACTGGTTGAACAGTTCAATCTGCCGATTGA
 GAAAACGGTTAAGACTCTGCTGGTTAAAGCGGTTGAAGGCAGCAGCTTCCCG
 CAGGTTGCGCTGCTGGTGCAGCGGTGATCACGAGCTGAACGAAGTTAAAGCAG
 AAAAAGTGCCGCAGGTTGCAAGCCCGCTGACTTTCGCGACCGAAGAAGAAAT
 TCGTGCCGTGGTTAAAGCCGGTCCGGGTTCACTGGGTCCGGTAAACATGCCG
 ATTCCGGTGGTGATTGACCGTACCGTTGCGGCGATGAGTGATTTGCTGCTGG
 TGCTAACATCGATGGTAAACACTACTTCGGCATCAACTGGGATCGCGATGTC
 GCTACCCCGGAAGTTGCAGATATCCGTAACGTGGTGGCTGGCGATCCAAGCC
 CGGATGGCCAGGGTAGGCTGCTGATCAAACGTGGTATCGAAGTTGGTCACAT
 CTTCCAGCTGGGTACCAAGTACTCCGAAGCACTGAAAGCCTCCGTACAGGGT
 GAAGATGGCCGTAACCAAATCCTGACGATGGGTTGCTACGGTATCGGGGTAA
 CGCGTGTGGTAGCTGCGGCGATTGAGCAGAAGTACGACGAACGAGGCATCGT

ATGGCCTGACGCTATCGCGCCGTTCCAGGTGGCGATTCTGCCGATGAACATG
 CACAAATCCTTCCGCGTACAAGAGCTTGCTGAGAACTGTACAGCGAACTGC
 GTGCACAAGGTATCGAAGTGCTGCTGGATGACCGCAAAGAGCGTCCGGGCGT
 GATGTTTGCTGATATGGAAGTATCGGTATTCCGCACACTATTGTGCTGGGCG
 ACCGTAACCTCGACAACGACGATATCGAATATAAATATCGTCGCAACGGCGA
 GAAACAGTTAATTAAGACTGGTGACATCGTTCGAATATCTGGTGAAACAGATT
 AAAGGCTGA

MRTSQYLLSTLKETPADAEVISHQLMLRAGMIRKGLASGLYTWLPTGVRVLKKVE
 NIVREEMNAGAIEVSMPVVQPADLWQESGRWEQYGPPELLRFVDRGERPFVLG
 PTHEEVITDLIRNELSSYKQLPLNFYQIQTKFRDEVPRFGVMRSREFLMKDAYSF
 HTSQESLQETYDAMYAAYSKIFSRMGLDFRAVQADTGSIGGSASHEFQVLAQSG
 EDDVVFSDTSDYAANIELAEAIAPKEPRAAATQEMTLVDTPNAKTIAELVEQFNL
 PIEKTVKTLVKAVEGSSFPQVALLVRGDHELNEVKAELPQVASPLTFATEEEIR
 AVVKAGPGSLGPVNMPIPVVIDRTVAAMSDFAAAGANIDGKHYFGINWDRDVAT
 PEVADIRNVVAGDPSPDGQGRLLIKRGIEVGHIFQLGTYSEALKASVQGEDGRN
 QILTMGCYGIGVTRVVAIAEQNYDERGIVWPDAIAPFQVAILPMNMHKSFRVQ
 ELAEKLYSELRAQGIEVLLDDRKERPGVMFADMELIGIPHTIVLGDRNLDNDIE
 YKYRRNGEKQLIKTGDIVEYLVKQIKG

2.6.6 Screening *ncPro* incorporation

A single colony of *E. coli* strain CAG18515 harboring plasmid pQE80_H27R-PI_proS was used to inoculate a culture of Luria Bertani (LB) medium supplemented with ampicillin. The culture was grown overnight at 37°C until stationary phase was reached, then diluted 1:100 into 100 mL of 1x M9 medium, supplemented with all twenty amino acids. The medium composition of M9 is as follows: 8.5 mM NaCl, 18.7 mM NH₄Cl, 22 mM KH₂PO₄, 47.8 mM Na₂HPO₄, 0.1 mM CaCl₂, 1 mM MgSO₄, 3 mg L⁻¹ FeSO₄, 1 µg L⁻¹

trace metals [Cu^{2+} , Mn^{2+} , Zn^{2+} , MoO_4^{2-}], 35 mg L⁻¹ thiamine HCl, 10 mg L⁻¹ biotin, 20 mM D-glucose, 100 mg L⁻¹ ampicillin, 50 mg L⁻¹ of each L-amino acid.

The culture was grown at 37°C until it reached OD ~0.8, after which it was subjected to a medium shift: cells were pelleted via centrifugation (5 kg, 5 min, 4°C) and washed twice with 10 mL ice-cold 0.9% NaCl. Washed cells were resuspended in 80 mL of 1.25x M9 –Pro, a 1.25x concentrated form of M9 that omits proline. The culture was split into 4 mL aliquots, and incubated for 30 min at 37°C to deplete residual proline. A 1 mL solution containing 2.5 mM ncPro and 1.5 M NaCl was added (0.5 mM ncPro and 0.3 M NaCl working concentrations). After 30 min of incubation at 37°C to allow for ncPro uptake, proinsulin expression was induced by the addition of 1 mM IPTG. Cultures were incubated for 2.5 h at 37°C, after which cells were harvested via centrifugation and stored at -80°C until further processing.

Cell pellets were thawed and lysed with B-PER Complete (Thermo Fisher Scientific) for 1 h at room temperature with shaking, then centrifuged (20 kg, 10 min) and the supernatant discarded. The pellet (containing insoluble proinsulin) was washed once with Triton wash buffer (2 M urea, 20 mM Tris, 1% Triton X-100, pH 8.0), and twice with ddH₂O. The pellet was resuspended in solubilization buffer (8 M urea, 300 mM NaCl, 50 mM NaH₂PO₄, pH 8.0), and proinsulin was allowed to dissolve for 1 h at room temperature with shaking. Samples were centrifuged, and the supernatant removed for analysis by SDS-PAGE and MALDI-TOF (described in section 2.6.10 below).

2.6.7 Proinsulin expression

A single colony of *E. coli* strain CAG18515 harboring plasmid pEQ80_H27R-PI_proS (or the corresponding plasmid with a point mutation in the *proS* gene; see Table 4.S3) was used to inoculate 70 mL of LB medium containing ampicillin, and the culture was grown overnight at 37°C to stationary phase. The overnight culture was used to inoculate 5 L (as 4x1.25 L cultures) of 1x Andrew's Magical Medium (AMM),⁵¹ a defined medium containing all 20 proteinogenic amino acids, in 2.8 L Fernbach flasks. The composition of AMM was the following: 3.60 g L⁻¹ glucose, 3.5 g L⁻¹ KH₂PO₄, 6.56 g L⁻¹ K₂HPO₄•3H₂O, 3.5 g L⁻¹ (NH₄)₂HPO₄, 8.37 g L⁻¹ MOPS, 0.72 g L⁻¹ tricine, 2.92 g L⁻¹ NaCl, 0.51 g L⁻¹ NH₄Cl, 0.26 g L⁻¹ MgCl₂•7H₂O, 50 mg L⁻¹ K₂SO₄, 0.246 mg L⁻¹ MgSO₄•7H₂O, 12.3 mg L⁻¹ CaCl₂•2H₂O, 2.8 mg L⁻¹ FeSO₄•7H₂O, 0.5 mg L⁻¹ thiamine, 24 µg L⁻¹ boric acid, 1 µg L⁻¹ trace metals (Cu²⁺, Mn²⁺, Zn²⁺, MoO₄²⁻), and 50 mg L⁻¹ each amino acid.

When growth reached mid-exponential phase (OD₆₀₀ ~0.8), the culture was subjected to a medium shift: cells were pelleted via centrifugation (5 kg, 5 min, 4°C) and washed twice with 100 mL ice-cold 0.9% NaCl. Washed cells were resuspended in 1 L of 1.25x AMM –Pro, a 1.25x concentrated form of AMM that omits proline. Cells were incubated for 30 min at 37°C to deplete residual proline, after which 250 mL of a solution containing 2.5-5.0 mM ncPro (see Table 4.S3) and 2.5 M NaCl was added (0.5-1.0 mM ncPro and 0.5 M NaCl working concentrations). After 30 min of incubation at 37°C to allow for ncPro uptake, proinsulin expression was induced by the addition of 1 mM IPTG. Cultures were incubated overnight at 37°C, after which cells were harvested via centrifugation and stored at -80°C until further processing.

Proline-containing proinsulin was expressed using strain CAG18515 harboring plasmid pQE80-H27R-PI_proS in 7.5 L (as 6 x 1.25 L cultures) of Terrific Broth (TB). 1 mM IPTG was added at mid-log phase ($OD_{600} \sim 0.8$) to induce proinsulin expression. Cultures were incubated at 37°C for 3 h, after which cells were harvested via centrifugation and stored at -80°C until further processing.

2.6.8 Proinsulin refolding

Cell pellets were warmed from -80°C to room temperature and resuspended in 5 mL IB buffer (50 mM tris, 100 mM NaCl, 1 mM EDTA, pH 8.0) per gram cell pellet. 1 mg L⁻¹ lysozyme and 1 mM PMSF were added, and the slurry sat on ice for 30 min before cells were lysed via sonication. The lysate was centrifuged (14 kg, 30 min, 4°C) and the soluble fraction was discarded. The pellet was washed twice with IB buffer + 1% Triton X-100, once with IB buffer, and once with water; this final step required centrifugation for 45 min. The washed inclusion body pellet was resuspended in a minimal amount of water, and the mass of proinsulin in the inclusion body pellet was estimated by SDS-PAGE.

In preparation for proinsulin refolding, the inclusion body was resuspended in 3 M urea and 10 mM cysteine in water, such that the proinsulin concentration was 1 mg proinsulin per L total slurry. To dissolve proinsulin, the pH was adjusted to 12 and sample stirred for 1 h at room temperature. At this stage, ncPro incorporation was assessed by MALDI-TOF, which is described in section 2.6.10 below. The solubilized proinsulin solution was diluted ten-fold into refold buffer (10 mM CAPS, pH 10.6) that had been pre-cooled to 4°C. The pH of the refold solution was adjusted to 10.7 and the sample stored at 4°C; care was taken

to ensure that the solution pH remained between 10.6 and 10.8 throughout the refolding process. Proinsulin refolding progress was monitored by reverse-phase HPLC, and usually reached completion within 50 h.

Proinsulin was enriched from the refold solution after adjusting the pH to 8.0 and incubating the sample overnight with Ni-NTA resin and 10 mM imidazole. The resin was washed with wash buffer (25 mM imidazole in PBS, pH 8.0), and proinsulin was eluted with elution buffer (250 mM imidazole in PBS, pH 8.0). Fractions containing proinsulin were combined and extensively dialyzed against 10 mM sodium phosphate, pH 8.0.

2.6.9 Insulin maturation and purification

Refolded proinsulin was warmed to 37°C and digested with trypsin (20 U mL⁻¹) and carboxypeptidase-B (10 U mL⁻¹) at 37°C for 90 min to remove the N-terminal tag and C chain. Digestion was halted by adjusting the pH to ~3 with 6 N HCl.

Insulins were immediately purified after proteolysis by reverse-phase HPLC on a C₄ column (Penomenex Jupiter 5 µm particle size, 300 Å pore size, 250x10 mm) using 0.1% TFA in water (solvent A) and 0.1% TFA in acetonitrile (solvent B) as mobile phases. A gradient of 25-32% solvent B was applied over 65 min, and fractions containing insulin were collected. Samples for purity analysis were removed at this stage; the remaining portion of the fraction was lyophilized. Each insulin fraction was analyzed by analytical reverse-phase HPLC, MALDI-TOF MS (Figure 2.2e-h), and SDS-PAGE (Figure 2.S2) to verify sample quality and ensure ≥95% purity for all downstream analyses. Lyophilized insulin powders were stored at -20°C until further use.

2.6.10 MALDI-TOF MS

To assess levels of incorporation of ncPros into the corresponding proinsulins, samples were digested with Glu-C, which results in a peptide fragment containing ProB28 (⁵⁰RGFFYTPKTRRE). A 20 μ L aliquot of proinsulin was subjected to cysteine reduction (5 mM DTT, 55°C for 20 min) and alkylation (15 mM iodoacetimide, RT for 15 min in the dark), prior to 10-fold dilution into 100 mM NH_4HCO_3 , pH 8.0 (100 μ L final volume). Digestion was started with addition of 0.6 μ L Glu-C (0.5 $\mu\text{g } \mu\text{L}^{-1}$ in ddH₂O) at 37°C for 2.5 h. The digestion reaction was quenched by adding 10 μ L of 5% TFA. Peptides were desalted using ZipTip C₁₈ columns (MilliporeSigma) according to the manufacturer's protocol. Desalted peptides (in 50% acetonitrile, 0.1% TFA) were diluted 3:1 into the matrix solution (α -cyanohydroxycinnamic acid in 50% ACN, 0.1% TFA) and analyzed by MALDI-TOF MS. Analog incorporation was calculated by comparing the area under the curve (AUC) of the ncPro form of the peptide ($m/z = 1572$ for 4*R*-Me and 4*S*-Me, and 1570 for 4ene) with the AUC of the canonical proline peptide ($m/z = 1558$).

HPLC-purified insulins were analyzed as full-length, mature proteins. Aliquots directly from HPLC purification (~30% ACN, 0.1% TFA) were mixed 1:1 with matrix solution (sinapic acid in 30% ACN, 0.1% TFA) before analysis by MALDI-TOF MS.

2.6.11 Circular dichroism spectroscopy

Equilibrium measurements: The circular dichroism spectra of insulin samples (60 μ M in 100 mM sodium phosphate, pH 8.0) were measured at 25°C in 1 mm quartz cuvettes on an Aviv Model 430 Circular Dichroism Spectrophotometer using a step size of 0.5 nm and

averaging time of 1 s. A reference buffer spectrum was subtracted from each sample spectrum.

Kinetic measurements: Insulin samples in 100 mM sodium phosphate buffer pH 8.0 were dialyzed overnight against 28.6 mM tris buffer, pH 8.0 (Slide-A-Lyzer dialysis cassettes, 3.5 kDa MWCO, ThermoFisher). Insulins were formulated as the following: 600 μ M insulin, 250 μ M ZnCl₂, 25 mM *m*-cresol, 25 mM tris buffer, pH 8. To a stirred buffer solution containing 2.98 mL of 25 mM tris, pH 8.0 in a 10 mm quartz cuvette was injected 20 μ L of the insulin formulation (150-fold dilution). Ellipticity was monitored at 222 nm over 120 s (1 s kinetic interval, 0.5 s time constant, 10 nm bandwidth) at 25°C. A typical run led to a rapid drop in CD signal as mixing occurred (~5 s), then a gradual rise to an equilibrium ellipticity representative of an insulin monomer. Data preceding the timepoint with the greatest negative ellipticity (representing the mixing time) were omitted from further analysis. Runs were discarded if the maximum change in mean residue ellipticity from equilibrium did not exceed 750 deg cm² dmol⁻¹, which indicated poor mixing. The remaining data were fit to a mono-exponential function using Scipy (Python). The data presented here are from at least two separate insulin HPLC fractions, measured on two different days.

For quality control, an equilibrium spectrum for each protein was obtained after dilution as described above; all spectra resembled that of the insulin monomer²⁹ (data not shown). The CD spectrum of human insulin under pre-dilution formulation conditions was obtained

using a 0.1 mm quartz cuvette. In each case, a blank spectrum containing all buffers and ligands was subtracted from the sample spectrum.

2.6.12 Reduction of blood glucose in diabetic mice

NODscid mice (NOD.CB17-Prkdcscid/J) were obtained from Jax Mice (Bar Harbor, ME). Mice were maintained under specific pathogen-free conditions, and experiments were conducted according to procedures approved by the Institutional Animal Care and Use Committee (IACUC) at the City of Hope. Adult (8-12 week old) male NODscid mice were injected intraperitoneally ($45 \text{ mg kg}^{-1} \text{ day}^{-1}$ for 3 days) with freshly prepared streptozotocin (STZ) in 50 mM citrate buffer, pH 4.5 to induce diabetes. Diabetes was confirmed 3 weeks after the last dose of STZ by detection of high glucose levels ($200\text{-}600 \text{ mg dL}^{-1}$) as measured by a glucomonitor (Freestyle, Abbott Diabetes Care, Alameda, CA) in blood sampled from the lateral tail vein. Insulin analogs were diluted to $100 \text{ } \mu\text{g mL}^{-1}$ in formulation buffer (1.6 mg mL^{-1} *m*-cresol, 0.65 mg mL^{-1} phenol, 3.8 mg mL^{-1} sodium phosphate pH 7.4, 16 mg mL^{-1} glycerol, $0.8 \text{ } \mu\text{g mL}^{-1}$ ZnCl_2). Insulin analogs were injected ($35 \text{ } \mu\text{g kg}^{-1}$) subcutaneously at the scruff and blood glucose was measured at 0, 10, 20, 30, 40, 50, 60, 80, 100, 120, and 150 min.

2.6.13 Fibrillation

Insulin samples ($60 \text{ } \mu\text{M}$ in 100 mM sodium phosphate, pH 8.0) were centrifuged at 22,000 g for 1 h at 4°C , prior to the addition of $1 \text{ } \mu\text{M}$ thioflavin T (ThT). Each insulin ($200 \text{ } \mu\text{L}$) was added to a 96-well, black, clear bottom plate (Greiner Bio-One) and sealed. Samples were shaken continuously at 960 rpm on a Varioskan multimode plate reader at 37°C , and

fluorescence readings were recorded every 15 min (444 nm excitation, 485 nm emission). Fibrillation runs were performed on at least two separate HPLC fractions, each in triplicate or quadruplicate, on at least two different days. The growth phase of each fibrillation replicate was fit to a linear function, and fibrillation lag times were reported as the x-intercept of this fit. Fibril samples were stored at 4°C until analysis by TEM.

2.6.14 Transmission electron microscopy

Insulin fibrils were centrifuged (5 kg, 1 min), then washed twice and resuspended in ddH₂O. Fibrils were stained with 2% uranyl acetate on a 300-mesh formvar/carbon coated copper grid (Electron Microscopy Sciences) and imaged on a Tecnai T12 LaB6 120 eV transmission electron microscope.

2.6.15 ANS fluorescence

1 μM insulin or insulin variant was mixed with 5 μM ANS in 100 mM phosphate buffer, pH 8.0. Fluorescence emission spectra were measured in 1 cm quartz cuvettes at ambient temperature using a PTI QuantaMaster fluorescence spectrofluorometer. A 350 nm excitation wavelength and scan rate of 2 nm s⁻¹ were used. Measurements for each insulin were performed in triplicate from three separate HPLC fractions. Spectra were smoothed before plotting and determining the emission maxima.

2.6.16 Analytical ultracentrifugation

Insulins were dissolved in 100 mM phosphate buffer, pH 8.0 and formulated at 60 μM. Alternatively, insulins were dialyzed against 28.6 mM tris buffer, pH 8.0, and formulated at 300 μM insulin, 12.5 mM *m*-cresol, and 125 μM ZnCl₂. Ligand-free insulins in tris buffer

were formulated from the same dialysis sample. Velocity sedimentation experiments were performed by Amy Henrickson and Maduni Ranasinghe at the Canadian Center for Hydrodynamics at the University of Lethbridge, with input from Borries Demeler. Samples containing 12.5 mM *m*-cresol were measured by interference optics, due to the high background absorbance of *m*-cresol. Ligand-free samples were measured using absorbance optics. All data were analyzed with UltraScan III version 4.0 release 6606.⁵² Sedimentation coefficients were determined from the enhanced van Holde-Weischhet analysis⁵³ to generate diffusion-corrected sedimentation coefficient distributions.

2.6.18 Calculations of proline and proline analog conformation

The equilibrium geometry conformations of the N-methyl, O-methyl ester protected versions of proline, 4-methyleneproline, and 3,4-dehydropyridine in water were calculated using Spartan Student (Wavefunction) at the B3LYP/G-31+G** level of theory. Pseudorotation parameters were calculated from the dihedral angles about the pyrrolidine ring, as previously reported.⁵⁴

2.6.19 Models of ins-4R-Me and ins-4S-Me hexamers

Crystal structures of the T₆ (PDB: 1MSO) and R₆ (1EV6) insulin hexamers were downloaded from the Protein Data Bank and visualized with Pymol. The hydrogen atoms at the C^γ position of ProB28 were replaced with methyl groups; no additional energy minimization was used.

2.7 Supplementary figures and tables

Figure 2.S1. SDS-PAGE analysis of proinsulin expression in media supplemented with non-canonical proline analogs. Proinsulin (12.7 kDa) was expressed after a medium shift to ncPro-containing medium. The inclusion body fraction was isolated, solubilized, and analyzed by SDS-PAGE.

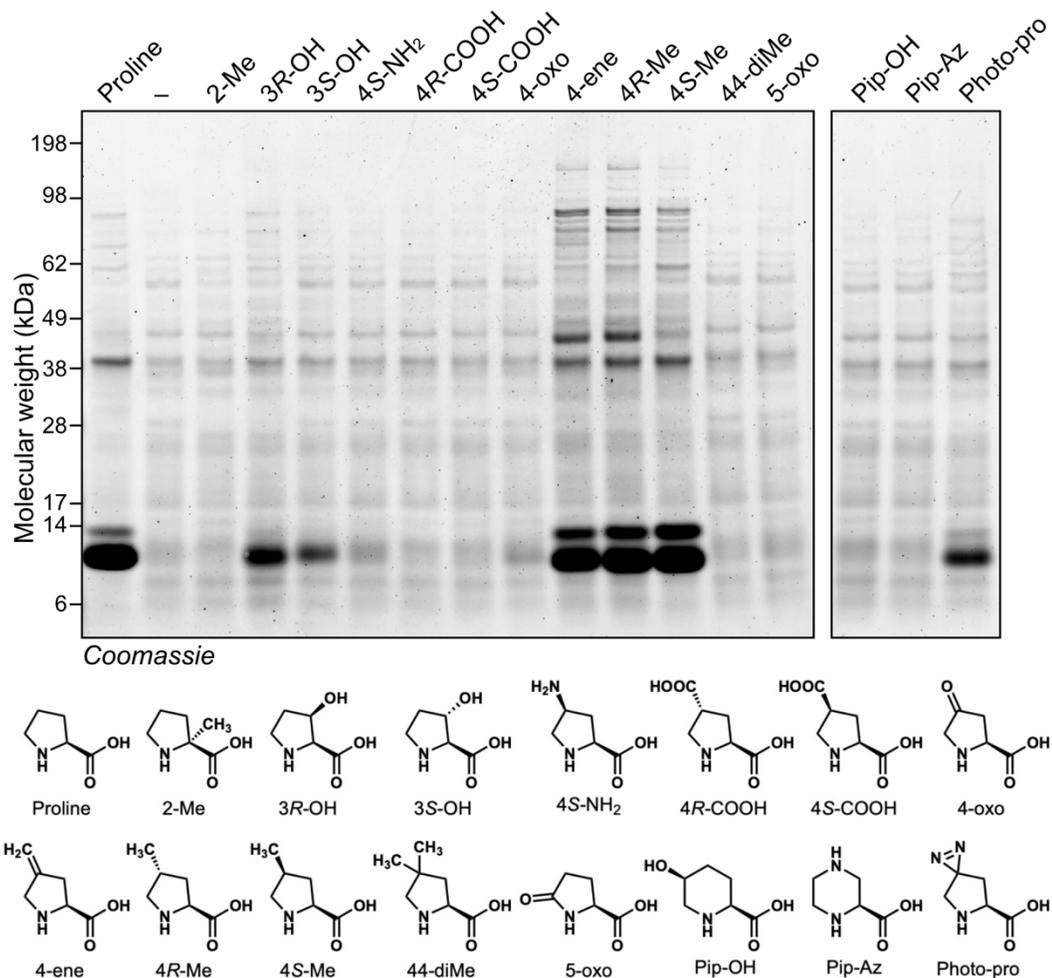


Figure 2.S2. Purity of insulin samples assessed by SDS-PAGE. HPLC-purified insulin (a), Ins-4*R*-Me (b), Ins-4*S*-Me (c), and Ins-4ene were analyzed by SDS-PAGE to validate purity; shown are representative lanes corresponding to individual HPLC fractions.

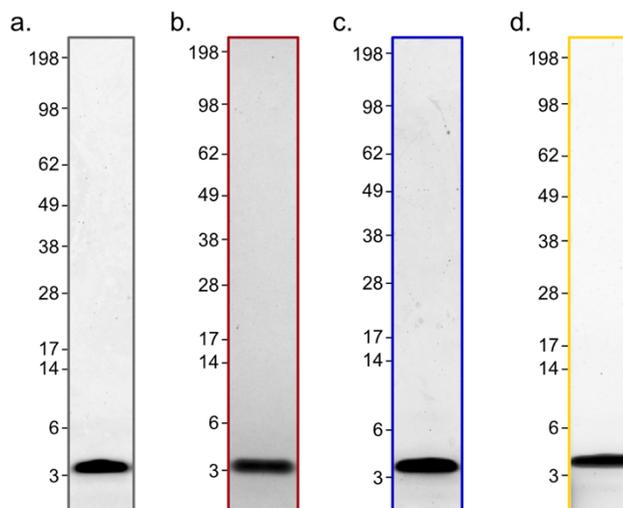


Figure 2.S3. Circular dichroism controls. At 60 μM , insulin is expected to exist as a dimer at pH 8, monomer in 20% ethanol, and denatured in the presence of 8 M guanidinium chloride. These spectra are overlaid with equilibrium spectra collected before and after dilution for kinetic CD measurements. Spectra below 210-215 nm were omitted for some samples due to high levels of buffer absorbance at these wavelengths.

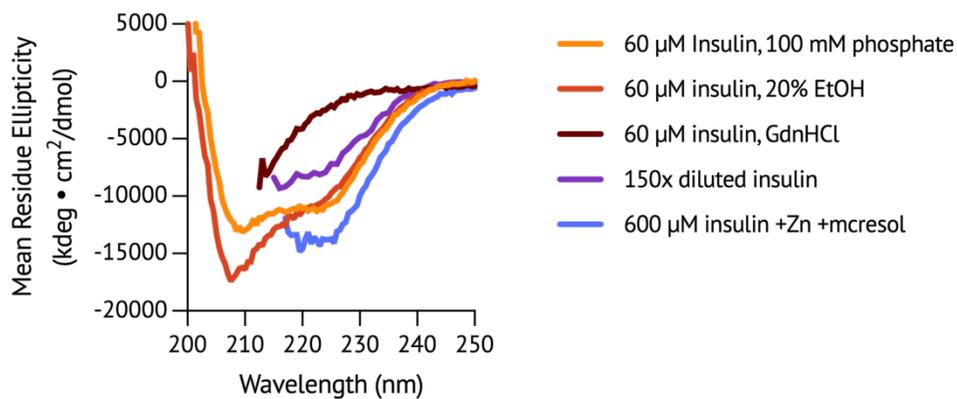


Figure 2.S4. Models of ins-4*R*-Me and ins-4*S*-Me hexamers. 4*S*-Me and 4*R*-Me were modeled in the structures of the R₆ (a) and T₆ (b) insulin hexamers (PDB ID: 1EV3 & 1MSO, respectively). Atoms near to each methyl substituent are indicated; distance measurements are in Å.

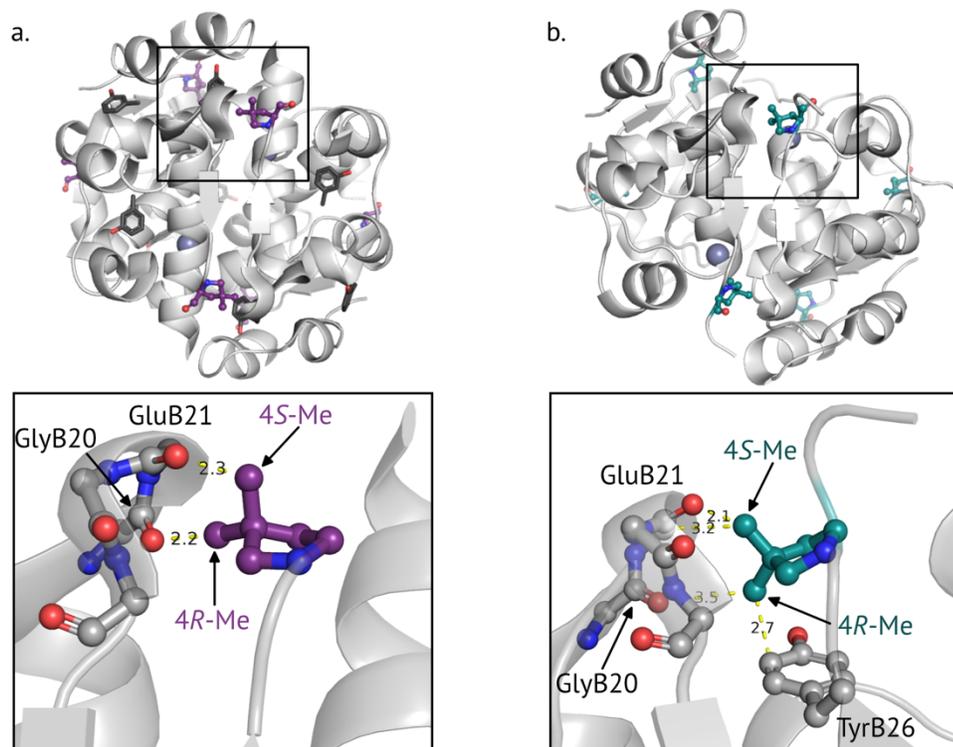


Figure 2.S5. Conformations of proline, 4-methyleneproline, and 3,4-dehydroproline.

The equilibrium geometry conformations of protected versions of proline (a), 4ene (b), and 3,4-dehydroproline (c) were calculated (B3LYP/G-31+G**). The pseudorotation parameters amplitude (A) and phase angle (P)⁵⁵ are indicated for each structure. Amplitude corresponds to the degree of puckering for each proline, and phase angle represents puckering geometry. The *endo* (P~198°) and *exo* (P~18°) ring puckers of proline are nearly isoenergetic and rapidly interconvert.⁵⁴ More notable in this case is the puckering amplitude, which decreases with the addition of sp² hybridized carbon atoms.

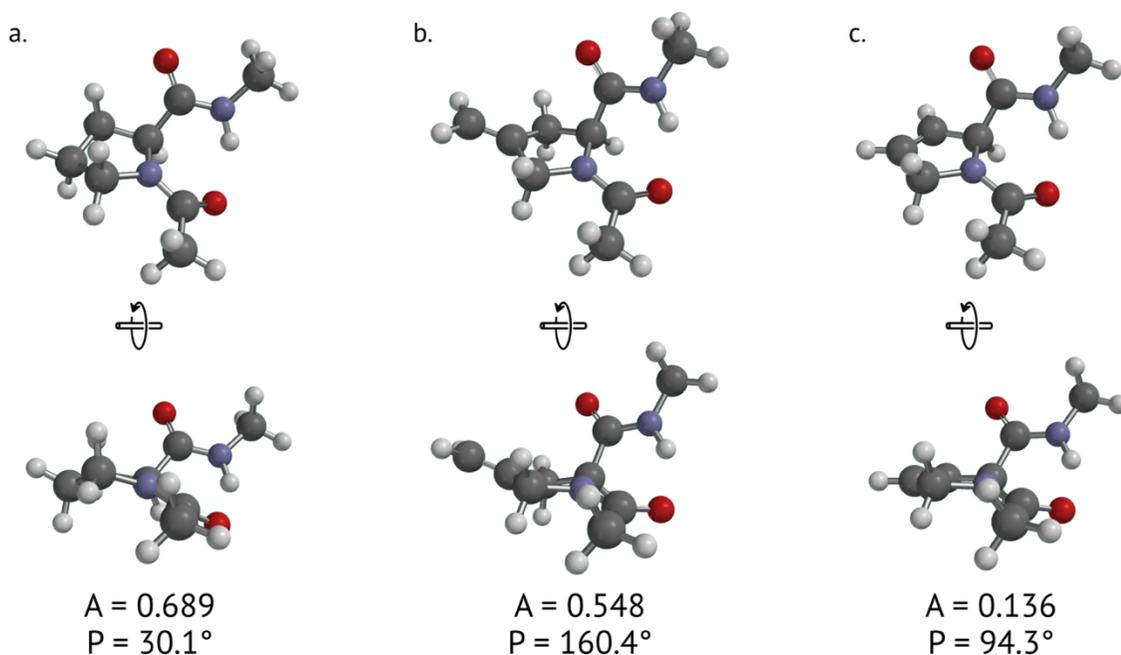


Table 2.S1. Incorporation of non-canonical proline residues

ncPro	Expected m/z	Observed m/z[‡]	Expected Δ m/z[#]	Observed Δ m/z^{‡#}	Incorporation efficiency[‡]
Proline	1557.90	1557.19 \pm 0.02	–	–	–
2-Me	1571.91	n.d.*	14.01	–	–
3 <i>R</i> -OH	1573.88	1573.184 \pm 0.006	15.98	15.5 \pm 0.7	0.672 \pm 0.004
3 <i>S</i> -OH	1573.88	1573.184 \pm 0.008	15.98	15.990 \pm 0.005	0.537 \pm 0.002
4 <i>S</i> -NH ₂	1572.90	1572.191 \pm 0.001	15.00	14.990 \pm 0.006	0.17 \pm 0.03
4 <i>R</i> -COOH	1601.89	n.d.	43.99	–	–
4 <i>S</i> -COOH	1601.89	n.d.	43.99	–	–
4-oxo	1571.87	1571.182 \pm 0.009	13.97	13.974 \pm 0.009	0.15 \pm 0.01
4-ene	1569.89	1569.218 \pm 0.008	11.99	11.971 \pm 0.001	0.926 \pm 0.002
4 <i>R</i> -Me	1571.91	1571.241 \pm 0.006	14.01	13.90 \pm 0.02	0.853 \pm 0.004
4 <i>S</i> -Me	1571.91	1571.26 \pm 0.01	14.01	14.017 \pm 0.004	0.775 \pm 0.007
44-diMe	1585.93	n.d.	28.03	–	–
5-oxo	1571.87	n.d.	13.97	–	–
Pip-OH	1587.91	n.d.	30.01	–	–
Pip-Az	1572.90	n.d.	15.00	–	–
Photo-pro [%]	528.61	528.61	8.66	8.67	0.373

[‡]Average \pm standard deviation of two technical replicates

[#]Mass shift compared to the proline-containing peptide present in the spectrum

* n.d., not detected

[%]Digested photo-proline peptide was analyzed by LC-ESI-MS, due to diazirine photolysis during MALDI-TOF analysis. We quantified the $[M+3H]^{+3}$ ion for the proline and ncPro-containing peptides. We also note the presence (27%; $m/z = 519.3$) of a 3,4-dehydroproline-containing peptide that likely results from photo-pro degradation within the culture, followed by co-translational incorporation (see Ch. 4). The incorporation efficiency for photo-pro reported here is calculated with respect to the proline and dehydroproline ions.

Table 2.S2. Mass spectrometry characterization of insulin variants

Protein	Digested peptide			Mature insulin	
	Expected m/z	Observed m/z [‡]	Incorporation efficiency [‡]	Expected m/z	Observed m/z [‡]
Insulin	1557.78	1557.43 ± 0.06	–	5808.6	5807.7 ± 0.4
Ins-4R-Me	1571.86	1571.83 ± 0.03	0.901 ± 0.010	5822.6	5822.2 ± 0.3
Ins-4S-Me	1571.86	1571.70 ± 0.03	0.896 ± 0.002	5822.6	5822.34 ± 0.01
Ins-4ene	1569.84	1569.56 ± 0.02	0.9417 ± 0.0005	5820.6	5821.2 ± 0.3

[‡]Average ± standard deviation of two technical replicates

Table 2.S3. Expression conditions and insulin yields

Protein	ProRS variant	[ncPro] (mM)	[NaCl] (M)	Approx. proinsulin yield (mg L ⁻¹) [‡]	Approx. mature insulin yield (mg L ⁻¹)
Insulin*	wt	–	–	35	8.0
Ins-4R-Me	C443G	0.5	0.5	29	3.9
Ins-4S-Me	wt	0.5	0.5	53	7.9
Ins-4ene	M157Q	1.0	0.5	23	1.5

[‡]Yields determined by measuring absorbance (280 nm) after proinsulin refolding and Ni-NTA enrichment.

*Expressed in terrific broth (TB)

Table 2.S4. Summary of insulin variant characterization

Protein	Ellipticity ratio (208/222 nm) [‡]	Sedimentation coefficient [‡]	Fibrillation lag time (h) [‡]	Hexamer dissociation t _{1/2} (s)	ANS emission maximum (nm) [*]
Insulin	1.24 ± 0.03	1.6S	16.6 ± 4.1	18.4 ± 2.8	470 ± 5
Ins-4R-Me	1.37 ± 0.05	1.3S	15.0 ± 3.5	9.8 ± 1.8	465 ± 3
Ins-4S-Me	1.19 ± 0.06	1.4S	12.5 ± 2.5	9.9 ± 2.5	466 ± 6
Ine-4ene	1.32 ± 0.05	1.3S	8.2 ± 4.0	17.0 ± 2.3	450 ± 11

[‡]60 μM insulin, 100 mM phosphate buffer, pH 8.0

^{*}1 μM insulin, 5 μM ANS, 100 mM phosphate buffer, pH 8.0

Table 2.S5. Sedimentation coefficients of insulin and ins-4S-Me

Protein	Formulation[*]	Dilution[‡]	Ligand-free[%]
Insulin	3.3S	1.1S	2.4S
Ins-4S-Me	3.4S	1.1S	3.1S

^{*}300 μ M insulin, 125 μ M ZnCl₂, 12.5 mM *m*-cresol, 25 mM tris, pH 8.0

[‡]4 μ M insulin, 1.7 μ M ZnCl₂, 167 μ M *m*-cresol, 25 mM tris, pH 8.0

[%]300 μ M insulin, 25 mM tris, pH 8.0

2.8 References

- (1) Shoulders, M. D.; Raines, R. T. Collagen Structure and Stability. *Annu Rev Biochem* **2009**, *78*, 929–958. <https://doi.org/10.1146/annurev.biochem.77.032207.120833>.
- (2) Ganguly, H. K.; Basu, G. Conformational Landscape of Substituted Prolines. *Biophys Rev* **2020**, *12*, 25–39. <https://doi.org/10.1007/s12551-020-00621-8>.
- (3) Lummis, S. C. R.; Beene, D. L.; Lee, L. W.; Lester, H. A.; Broadhurst, R. W.; Dougherty, D. A. *Cis–Trans* Isomerization at a Proline Opens the Pore of a Neurotransmitter-Gated Ion Channel. *Nature* **2005**, *438*, 248–252. <https://doi.org/10.1038/nature04130>.
- (4) Kim, W.; Hardcastle, K. I.; Conticello, V. P. Fluoroproline Flip-Flop: Regiochemical Reversal of a Stereoelectronic Effect on Peptide and Protein Structures. *Angew Chem Int Ed* **2006**, *45*, 1841–1845. <https://doi.org/10.1002/anie.200603227>.
- (5) Kim, W.; George, A.; Evans, M.; Conticello, V. P. Cotranslational Incorporation of a Structurally Diverse Series of Proline Analogues in an *Escherichia coli* Expression System. *ChemBioChem* **2004**, *5* (7), 928–936. <https://doi.org/10.1002/cbic.200400052>.
- (6) Kim, W.; McMillan, R. A.; Snyder, J. P.; Conticello, V. P. A Stereoelectronic Effect on Turn Formation Due to Proline Substitution in Elastin-Mimetic Polypeptides. *J Am Chem Soc* **2005**, *127* (51), 18121–18132. <https://doi.org/10.1021/ja054105j>.
- (7) Holmgren, S. K.; Taylor, K. M.; Bretscher, L. E.; Raines, R. T. Code for Collagen's Stability Deciphered. *Nature* **1998**, *392*, 666–667. <https://doi.org/10.1038/33573>.
- (8) Torbeev, V. Y.; Hilvert, D. Both the *Cis–Trans* Equilibrium and Isomerization Dynamics of a Single Proline Amide Modulate β 2-Microglobulin Amyloid Assembly. *Proc Natl Acad Sci* **2013**, *110* (50), 20051–20056. <https://doi.org/10.1073/pnas.1310414110>.
- (9) Torbeev, V.; Ebert, M. O.; Dolenc, J.; Hilvert, D. Substitution of Proline³² by α -Methylproline Preorganizes β 2-Microglobulin for Oligomerization but Not for Aggregation into Amyloids. *J Am Chem Soc* **2015**, *137* (7), 2524–2535. <https://doi.org/10.1021/ja510109p>.
- (10) Breunig, S. L.; Tirrell, D. A. Incorporation of Proline Analogs into Recombinant Proteins Expressed in *Escherichia coli*. *Methods Enzymol* **2021**, *656*, 545–571. <https://doi.org/10.1016/BS.MIE.2021.05.008>.
- (11) Haeusler, R. A.; McGraw, T. E.; Accili, D. Biochemical and Cellular Properties of Insulin Receptor Signalling. *Nat Rev Mol Cell Biol* **2018**, *19*, 31–44. <https://doi.org/10.1038/nrm.2017.89>.
- (12) Goeddel, D. V.; Kleid, D. G.; Bolivar, F.; Heyneker, H. L.; Yansura, D. G.; Crea, R.; Hirose, T.; Kraszewski, A.; Itakura, K.; Riggs, A. D. Expression in *Escherichia coli* of Chemically Synthesized Genes for Human Insulin. *Proc Natl Acad Sci* **1979**, *76* (1), 106–110. <https://doi.org/10.1073/pnas.76.1.106>.
- (13) Rakatzi, I.; Ramrath, S.; Ledwig, D.; Dransfeld, O.; Bartels, T.; Seipke, G.; Eckel, J. A Novel Insulin Analog With Unique Properties: LysB3 ,GluB29 Insulin Induces

- Prominent Activation of Insulin Receptor Substrate 2, but Marginal Phosphorylation of Insulin Receptor Substrate 1. *Diabetes* **2003**, *52*, 2227–2238. <https://doi.org/10.2337/diabetes.52.9.2227>.
- (14) Brange, J.; Ribel, U.; Hansen, J. F.; Dodson, G.; Hanse, M. T.; Havelund, S.; Melberg, S. G.; Norris, F.; Norris, K.; Snel, L.; Sørensen, A. R.; Voigt, H. O. Monomeric Insulins Obtained by Protein Engineering and Their Medical Implications. *Nature* **1988**, *333*, 679–682. <https://doi.org/10.1038/333679a0>.
- (15) Brems, D. N.; Alter, L. A.; Beckage, M. J.; Chance, R. E.; Dimarchi, R. D.; Green, L. K.; Long, H. B.; Pekar, A. H.; Shields, J. E.; Frank, B. H. Altering the Association Properties of Insulin by Amino Acid Replacement. *Protein Eng* **1992**, *5* (6), 527–533. <https://doi.org/10.1093/protein/5.6.527>.
- (16) Lepore, M.; Pampanelli, S.; Fanelli, C.; Porcellati, F.; Bartocci, L.; Di Vincenzo, A.; Cordoni, C.; Costa, E.; Brunetti, P.; Bolli, G. B. Pharmacokinetics and Pharmacodynamics of Subcutaneous Injection of Long-Acting Human Insulin Analog Glargine, NPH Insulin, and Ultralente Human Insulin and Continuous Subcutaneous Infusion of Insulin Lispro. *Diabetes* **2000**, *49*, 2142–2148. <https://doi.org/10.2337/diabetes.49.12.2142>.
- (17) Mathieu, C.; Gillard, P.; Benhalima, K. Insulin Analogues in Type 1 Diabetes Mellitus: Getting Better All the Time. *Nat Rev Endocrinol* **2017**, *13*, 385–399. <https://doi.org/10.1038/nrendo.2017.39>.
- (18) Bolli, G. B.; Di Marchi, R. D.; Park, G. D.; Pramming, S.; Koivisto, V. A. Insulin Analogues and Their Potential in the Management of Diabetes Mellitus. *Diabetologia* **1999**, *42* (10), 1151–1167. <https://doi.org/10.1007/s001250051286>.
- (19) Owens, D.; Vora, J. Insulin Aspart: A Review. *Expert Opin Drug Metab Toxicol* **2006**, *2* (5), 793–804. <https://doi.org/10.1517/17425255.2.5.793>.
- (20) Holleman, F.; Hoekstra, J. B. L. Insulin Lispro. *N Engl J Med* **1997**, *337* (3), 176–183. <https://doi.org/10.1056/NEJM199707173370307>.
- (21) Bakaysa, D. L.; Radziuk, J.; Havel, H. A.; Brader, M. L.; Li, S.; Dodd, S. W.; Beals, J. M.; Pekar, A. H.; Brems, D. N. Physicochemical Basis for the Rapid Time-Action of LysB28ProB29-Insulin: Dissociation of a Protein-Ligand Complex. *Protein Sci* **1996**, *5* (12), 2521–2531. <https://doi.org/10.1002/pro.5560051215>.
- (22) Brange, J.; Langkjaer, L. Insulin Structure and Stability. In *Stability and Characterization of Protein and Peptide Drugs: Case Histories*; Wang, Y. J., Pearlman, R., Eds.; Plenum Press: New York, 1993; pp 315–350. https://doi.org/10.1007/978-1-4899-1236-7_11.
- (23) Brange, J.; Havelund, S.; Hougaard, P. Chemical Stability of Insulin. 2. Formation of Higher Molecular Weight Transformation Products During Storage of Pharmaceutical Preparations. *Pharm Res* **1992**, *9* (6), 727–734. <https://doi.org/10.1023/a:1015887001987>.
- (24) Brange, J.; Langkjær, L.; Havelund, S.; Vølund, A. Chemical Stability of Insulin. 1. Hydrolytic Degradation During Storage of Pharmaceutical Preparations. *Pharm Res* **1992**, *9* (6), 715–726. <https://doi.org/10.1023/a:1015835017916>.

- (25) Weiss, M. A. Design of Ultra-Stable Insulin Analogues for the Developing World. *J Health Specialties* **2013**, *1* (2), 59–70. <https://doi.org/10.4103/1658-600X.114683>.
- (26) Kerr, D.; Morton, J.; Whately-Smith, C.; Everett, J.; Begley, J. P. Laboratory-Based Non-Clinical Comparison of Occlusion Rates Using Three Rapid-Acting Insulin Analogs in Continuous Subcutaneous Insulin Infusion Catheters Using Low Flow Rates. *J Diabetes Sci Technol* **2008**, *2* (3), 450–455. <https://doi.org/10.1177/193229680800200314>.
- (27) Fang, K. Y.; Lieblisch, S. A.; Tirrell, D. A. Replacement of ProB28 by Pipecolic Acid Protects Insulin against Fibrillation and Slows Hexamer Dissociation. *J Polym Sci A Polym Chem* **2019**, *57* (3), 264–267. <https://doi.org/10.1002/pola.29225>.
- (28) Lieblisch, S. A.; Fang, K. Y.; Cahn, J. K. B.; Rawson, J.; LeBon, J.; Teresa Ku, H.; Tirrell, D. A. 4*S*-Hydroxylation of Insulin at ProB28 Accelerates Hexamer Dissociation and Delays Fibrillation. *J Am Chem Soc* **2017**, *139*, 8384–8387. <https://doi.org/10.1021/jacs.7b00794>.
- (29) Pocker, Y.; Biswas, S. B. Conformational Dynamics of Insulin in Solution. Circular Dichroic Studies. *Biochemistry* **1980**, *19* (22), 5043–5049. <https://doi.org/10.1021/bi00563a017>.
- (30) Antolíková, E.; Žáková, L.; Turkenburg, J. P.; Watson, C. J.; Hančlová, I.; Šanda, M.; Cooper, A.; Kraus, T.; Brzozowski, A. M.; Jiráček, J. Non-Equivalent Role of Inter- and Intramolecular Hydrogen Bonds in the Insulin Dimer Interface. *J Biol Chem* **2011**, *286* (42), 36968–36977. <https://doi.org/10.1074/jbc.M111.265249>.
- (31) Plum, A.; Agersø, H.; Andersen, L. Pharmacokinetics of the Rapid-Acting Insulin Analog, Insulin Aspart, in Rats, Dogs, Pigs, and Pharmacodynamics of Insulin Aspart in Pigs. *Drug Metab Dispos* **2000**, *28*, 155–160.
- (32) Scapin, G.; Dandey, V. P.; Zhang, Z.; Prorise, W.; Hruza, A.; Kelly, T.; Mayhood, T.; Strickland, C.; Potter, C. S.; Carragher, B. Structure of the Insulin Receptor-Insulin Complex by Single-Particle Cryo-EM Analysis. *Nature* **2018**, *556* (7699), 122–125. <https://doi.org/10.1038/nature26153>.
- (33) Menting, J. G.; Whittaker, J.; Margets, M. B.; Whittaker, L. J.; Kong, G. K.-W.; Smith, B. J.; Watson, C. J.; Žáková, L.; Kletvíková, E.; Jiracek, J.; Chan, S. J.; Steiner, D. F.; Dodson, G. G.; Brzozowski, A. M.; Weiss, M. A.; Ward, C. W.; Lawrence, M. C. How Insulin Engages Its Primary Binding Site on the Insulin Receptor. *Nature* **2013**, *493*, 241–245. <https://doi.org/10.1038/nature11781>.
- (34) Jansen, R.; Dzwolak, W.; Winter, R. Amyloidogenic Self-Assembly of Insulin Aggregates Probed by High Resolution Atomic Force Microscopy. *Biophys J* **2005**, *88*, 1344–1353. <https://doi.org/10.1529/biophysj.104.048843>.
- (35) Nielsen, L.; Frokjaer, S.; Brange, J.; Uversky, V. N.; Fink, A. L. Probing the Mechanism of Insulin Fibril Formation with Insulin Mutants. *Biochemistry* **2001**, *40* (28), 8397–8409. <https://doi.org/10.1021/bi0105983>.

- (36) Jiménez, J. L.; Nettleton, E. J.; Bouchard, M.; Robinson, C. V.; Dobson, C. M.; Saibil, H. R.; Caspar, D. L. D. The Protofilament Structure of Insulin Amyloid Fibrils. *Proc Natl Acad Sci* **2002**, *99* (14), 9196–9201. <https://doi.org/10.1073/pnas.1424599399>.
- (37) Akbarian, M.; Yousefi, R.; Farjadian, F.; Uversky, V. N. Insulin Fibrillation: Toward Strategies for Attenuating the Process. *Chem Commun* **2020**, *56* (77), 11354–11373. <https://doi.org/10.1039/d0cc05171c>.
- (38) Semisotnov, G. V.; Rodionova, N. A.; Razgulyaev, O. I.; Uversky, V. N.; Gripas', A. F.; Gilmanshin, R. I. Study of the “Molten Globule” Intermediate State in Protein Folding by a Hydrophobic Fluorescent Probe. *Biopolymers* **1991**, *31* (1), 119–128. <https://doi.org/10.1002/BIP.360310111>.
- (39) Goldman, J.; Carpenter, F. H. Zinc Binding, Circular Dichroism, and Equilibrium Sedimentation Studies on Insulin (Bovine) and Several of Its Derivatives. *Biochemistry* **1963**, *13* (22), 4566–4574. <https://doi.org/10.1021/bi00719a015>.
- (40) Gast, K.; Schüler, A.; Wolff, M.; Thalhammer, A.; Berchtold, H.; Nagel, N.; Lenherr, G.; Hauck, G.; Seckler, R. Rapid-Acting and Human Insulins: Hexamer Dissociation Kinetics upon Dilution of the Pharmaceutical Formulation. *Pharm Res* **2017**, *34* (11), 2270–2286. <https://doi.org/10.1007/s11095-017-2233-0>.
- (41) Smith, G. D.; Ciszak, E.; Magrum, L. A.; Pangborn, W. A.; Blessing, R. H. R₆ Hexameric Insulin Complexed with *m*-Cresol or Resorcinol. *Acta Cryst* **2000**, *56* (12), 1541–1548. <https://doi.org/10.1107/S0907444900012749>.
- (42) Richards, J. P.; Stickelmeyer, M. P.; Flora, D. B.; Chance, R. E.; Frank, B. H.; DeFelippis, M. R. Self-Association Properties of Monomeric Insulin Analogs under Formulation Conditions. *Pharm Res* **1998**, *15* (9), 1434–1441. <https://doi.org/10.1023/a:1011961923870>.
- (43) Pekar, A. H.; Frank, B. H. Conformation of Proinsulin. A Comparison of Insulin and Proinsulin Self-Association at Neutral PH. *Biochemistry* **1972**, *11* (22), 4013–4016. <https://doi.org/10.1021/bi00772a001>.
- (44) Caporale, A.; Loughlin, J. O.; Ortin, Y.; Rubini, M. A Convenient Synthetic Route to (2*S*,4*S*)-Methylproline and Its Exploration for Protein Engineering of Thioredoxin. *Org Biomol Chem* **2022**. <https://doi.org/10.1039/D2OB01011A>.
- (45) Shoulders, M. D.; Hodges, J. A.; Raines, R. T. Reciprocity of Steric and Stereoelectronic Effects in the Collagen Triple Helix. *J Am Chem Soc* **2006**, *128*, 8112–8113. <https://doi.org/10.1021/ja061793d>.
- (46) Kang, Y. K.; Park, H. S. Conformational Preferences and *Cis–Trans* Isomerization of L-3,4-Dehydroproline Residue. *Peptide Sci* **2009**, *92* (5), 387–398. <https://doi.org/10.1002/BIP.21203>.
- (47) Loosli, S.; Foletti, C.; Pappmeyer, M.; Zürich, E. Synthesis of 4-(Arylmethyl)Proline Derivatives. *Synlett* **2019**, *30*, 508–510. <https://doi.org/10.1055/s-0037-1611672>.
- (48) Deming, T. J.; Fournier, M. J.; Mason, T. L.; Tirrell, D. A. Biosynthetic Incorporation and Chemical Modification of Alkene Functionality in Genetically Engineered

- Polymers. *J Macromol Sci A* **1997**, *34* (10), 2143–2150.
<https://doi.org/10.1080/10601329708010331>.
- (49) Bakaysa, D. L.; Radziuk, J.; Havel, H. A.; Brader, M. L.; Li, S.; Dodd, S. W.; Beals, J. M.; Pekar, A. H.; Brems, D. N. Physicochemical Basis for the Rapid Time-Action of LysB28ProB29-Insulin: Dissociation of a Protein-Ligand Complex. *Protein Sci* **1996**, *5*, 2521–2531.
- (50) Min, C. K.; Son, Y. J.; Kim, C. K.; Park, S. J.; Lee, J. W. Increased Expression, Folding and Enzyme Reaction Rate of Recombinant Human Insulin by Selecting Appropriate Leader Peptide. *J Biotechnol* **2011**, *151* (4), 350–356.
<https://doi.org/10.1016/j.jbiotec.2010.12.023>.
- (51) He, W.; Fu, L.; Li, G.; Andrew Jones, J.; Linhardt, R. J.; Koffas, M. Production of Chondroitin in Metabolically Engineered *E. coli*. *Metab Eng* **2015**, *27*, 92–100.
<https://doi.org/10.1016/J.YMBEN.2014.11.003>.
- (52) Demeler, B.; Gorbet, G. E. Analytical Ultracentrifugation Data Analysis with Ultrascan-III. *Analytical Ultracentrifugation: Instrumentation, Software, and Applications* **2016**, 119–143.
https://doi.org/10.1007/978-4-431-55985-6_8.
- (53) Demeler, B.; Van Holde, K. E. Sedimentation Velocity Analysis of Highly Heterogeneous Systems. *Anal Biochem* **2004**, *335* (2), 279–288.
<https://doi.org/10.1016/J.AB.2004.08.039>.
- (54) Park, H. S.; Kang, Y. K. Puckering Transition of the Proline Residue along the Pseudorotational Path: Revisited. *New J Chem* **2021**, *45* (22), 9780–9793.
<https://doi.org/10.1039/D1NJ01361K>.
- (55) Hudáky, I.; Baldoni, H. A.; Perczel, A. Peptide Models XXXVIII. Proline Conformers from X-Ray Crystallographic Database and from Ab Initio Computations. *J Mol Struct Theochem* **2002**, *582* (1–3), 233–249. [https://doi.org/10.1016/S0166-1280\(01\)00796-5](https://doi.org/10.1016/S0166-1280(01)00796-5).

*Chapter III***4S-FLUORINATION AT PROB29 OF INSULIN LISPRO SLOWS FIBRIL FORMATION****3.1 Contributions**

Chad D. Paavola, Michael P. Akers, Julie S. Moyers, and Alborz Mahdavi provided advice on proinsulin refolding. Alex M. Chapman prepared lispro variants used for mouse experiments. Janine Quijano and Jeanne LeBon performed mouse experiments, with input from Hsun Teresa Ku. Analytical ultracentrifugation was performed by Maduni Ranasinghe at the Canadian Center for Hydrodynamics at the University of Lethbridge, with input from Borries Demeler. Jay Winkler, Songye Chen, Scott Virgil, and Mona Shahgholi assisted with circular dichroism spectroscopy, transmission electron microscopy, chromatography, and mass spectrometry, respectively. Stephanie L. Breunig prepared all other insulin samples and performed all other experiments. Alex M. Chapman and David A. Tirrell contributed to the writing of this text.

3.2 Abstract

Recombinant insulin is a life-saving therapeutic for millions of patients affected by diabetes mellitus. Standard mutagenesis approaches have led to insulin variants that enable improved control of blood glucose: for instance, the fast-acting insulin lispro contains two point mutations that destabilize dimer formation and expedite absorption. However, insulins undergo irreversible denaturation upon storage, a process accelerated for the insulin monomer. Here we replace ProB29 of insulin lispro with 4*R*-fluoroproline, 4*S*-

fluoroproline, and 4,4-difloroproline. For most properties measured, fluorination at B29 does not change the behavior of lispro: all of the resulting variants reduce blood glucose in diabetic mice, sediment as monomers by analytical ultracentrifugation, exhibit similar secondary structure as measured by circular dichroism, and rapidly dissociate from the zinc and resorcinol-bound hexamer upon dilution. Though we measured similar lag times to fibril formation for most variants, we find that 4*S*-fluorination of ProB29 delays fibril formation. These results demonstrate how subtle molecular changes enabled by non-canonical amino acid mutagenesis can improve the stability of protein therapeutics.

3.3 Main text

Insulin, a 5.8 kDa peptide hormone composed of two disulfide-linked chains (A and B), is a widely-used, essential therapeutic for individuals with diabetes mellitus. Insulin is normally secreted from pancreatic β -cells in response to elevated blood glucose; its binding to the insulin receptor induces a variety of intracellular responses that lead to decreased blood glucose concentration.¹ Diabetes, which affects more than 400 million adults worldwide,² arises from dysfunction of insulin signaling, either from loss of insulin secretion (type 1), or from development of insulin resistance (type 2). Advances in insulin technology over the last century, notably the recombinant production of insulin,³ have dramatically reduced the mortality rate associated with diabetes. Individuals with type 1 diabetes typically rely on insulin replacement therapy to compensate for loss of pancreatic function, while those with type 2 often supplement lifestyle changes with oral medications that enhance insulin signaling, and in severe cases, receive exogenous insulin.

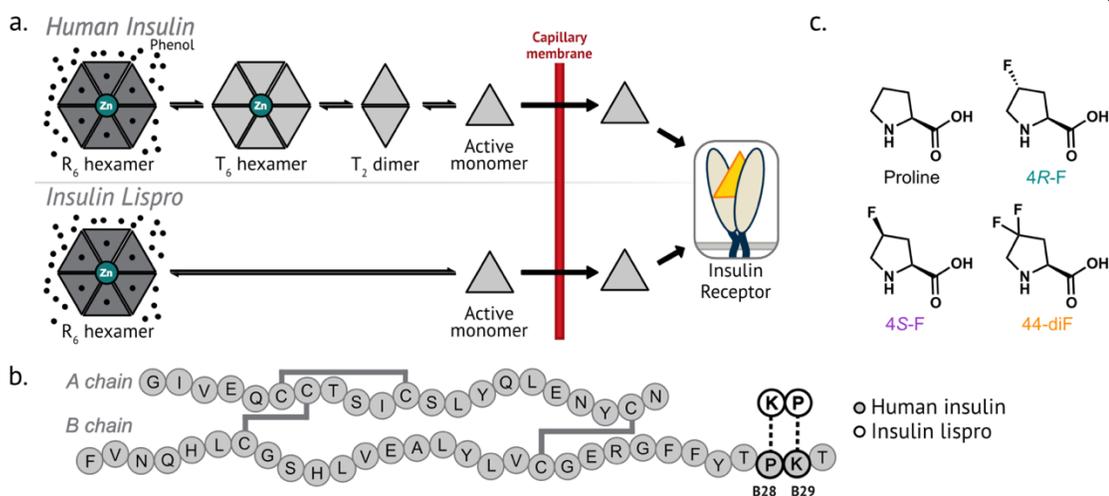


Figure 3.1. Proline mutagenesis at position B29 of insulin lispro. **a.** Simplified representation of the oligomerization states of human insulin and insulin lispro. Disrupted oligomerization speeds the release of lispro into the bloodstream. **b.** The amino acid sequences of human insulin and insulin lispro. **c.** The structures of proline and the 4-fluorinated proline analogs used in this study.

In its pharmaceutical formulations, insulin exists as a zinc- and phenol-bound hexamer (referred to as the R_6 state).⁴ Upon subcutaneous injection, insulin hexamers dissociate through lower-order oligomeric species before crossing the capillary membrane in the active monomeric state^{5,6} (Figure 3.1a). Because dissociation to the monomer is rate-limiting for insulin absorption into the bloodstream, disrupting insulin association also accelerates its onset of action.⁷ Introducing point mutations that destabilize oligomer formation has resulted in a suite of FDA-approved fast-acting insulins (FAIs).^{8–10} FAIs are rapidly released into the bloodstream shortly after injection,¹¹ resembling the transient increases in insulin concentration stimulated by elevated blood glucose (i.e., after a meal). Typical insulin replacement therapies combine regular treatments of basal (long-acting) insulin variants with injections of FAIs before mealtime to approximate the insulin-action

profile of a healthy pancreas.¹¹ However, FAIs tend to be more prone to chemical and physical denaturation, since the protective effects of oligomerization are interrupted.¹²

Insulin lispro (Humalog™, Eli Lilly), the first commercially-available FAI, disfavors monomer association through the inversion of two residues near the C-terminus of the insulin B-chain, ProB28 and LysB29 (Figure 3.1b).⁸ Removal of the constrained pyrrolidine ring of proline at position B28 grants additional conformational degrees of freedom and removes hydrophobic packing interactions at the dimer interface, disrupting the homodimer interaction.⁸ Consequently, insulin lispro hexamers dissociate to bioactive monomers considerably faster than human insulin, and are thus absorbed more rapidly.⁷ However, the standard mutagenesis approach employed in the discovery of insulin lispro⁸ (and other FDA-approved insulin variants^{9,10,13}) is limited by the chemical functionalities present in the 20 canonical amino acids, and restricted with respect to the chemical space that can be explored. In particular, no other canonical amino acid can recapitulate the unique structural and conformational properties of proline.¹⁴ In contrast, protein engineering strategies that draw upon non-canonical amino acids (ncAAs) afford access to chemistries beyond those found in natural or conventionally-engineered proteins.¹⁵

Non-canonical proline (ncPro) residues provide chemical diversity while maintaining proline's cyclic structure. Many ncPro residues exhibit conformational biases different from those of proline,¹⁴ so can be used to interrogate importance of proline conformation in protein behavior. ncPro residues with 4*R* electron-withdrawing groups (such as 4*R*-fluoroproline, 4*R*-F; Figure 3.1c) promote a *C^γ-exo* ring pucker due to a gauche effect,

which in turn enforces the *trans* amide isomer. Conversely, ncPro analogs with 4*S*-electron withdrawing groups (such as 4*S*-fluoroproline, 4*S*-F) favor the *C^γ-endo* pucker and the *cis* amide isomer, compared to canonical proline.¹⁴ The position of the amide *cis-trans* equilibrium preferences of 4,4-difluoroproline (44-diF) is similar to that of canonical proline.¹⁶ Isomerization of the 4-fluorinated prolines is accelerated relative to proline, as an inductive effect reduces the bond order of the preceding amide.¹⁶ Proline analogs have been used to determine the stereoelectronic origin of collagen stability,¹⁷ and have helped identify a key *cis-trans* isomerization event in the opening of the 5-HT₃ receptor.¹⁸

The consequences of replacing proline with fluorinated analogs depend on protein context. Replacing proline with ncPro residues in barstar,¹⁶ ubiquitin,¹⁹ and the trp cage miniprotein²⁰ stabilized the protein if the ncPro conformational preferences matched those present in proline; proline analogs with opposed preferences destabilized the protein. In the context of thioredoxin, the conformational preferences of the analog made little difference; replacing a conserved *cis* proline residue with either diastereomer of 4-fluoroproline led to similar effects on protein stability, despite the opposed conformational preferences of the fluoroprolines.²¹

A key proline residue at position 32 of β 2 microglobulin (β 2m) was replaced by the fluoroprolines 4*R*-F, 4*S*-F, and 44-diF,²² and by α -methylproline, which strongly favors the *trans* isomer.²³ Pro32 adopts the *cis* form in the natively folded protein, and the *trans* conformer in the amyloid fibril, suggesting that *cis-trans* isomerization is a key step in β 2m fibrillation. Consistent with this expectation, ncPro *cis-trans* preferences affected soluble

β 2m dynamics, oligomerization, and stability against thermal and chemical denaturation. However, the diverse fibrillation behaviors and fibril morphologies of the β 2m variants suggested that fibril formation is more complex than simple association of *trans*-Pro32 β 2m monomers.^{22,23}

We recently demonstrated that introduction of ncPro residues at position B28 in human insulin can be used to tune its biophysical properties,^{24,25} including results presented in Chapter II of this thesis. These ncPro mutagenesis experiments on human insulin prompted us to ask whether similar changes in the biophysical properties of insulin lispro might be achieved. To explore this question, we replaced ProB29 in insulin lispro with 4-fluorinated proline analogs (Figure 3.1c); these lispro variants will be referred to as KP-4*R*-F, KP-4*S*-F, and KP-44-diF.

Lispro variants were expressed in *Escherichia coli* as the corresponding proinsulin (a precursor to insulin lispro) under conditions that favor ncPro incorporation.²⁶ Briefly, the proline auxotrophic strain CAG18515 overexpressing the prolyl-tRNA synthetase was grown in M9 medium supplemented with all 20 amino acids until late log phase. Cells were washed and resuspended in a medium lacking proline, at which point high concentrations (0.3-0.5 M) of sodium chloride (to facilitate proline uptake), and the proline analog were added. Proline replacement was assessed by MALDI-TOF mass spectrometry of a proinsulin peptide fragment (Figure 3.2a-d), and of mature insulin (Figure 3.2e-h). Incorporation of all fluorinated proline analogs exceeded 90% (Table 3.S1).

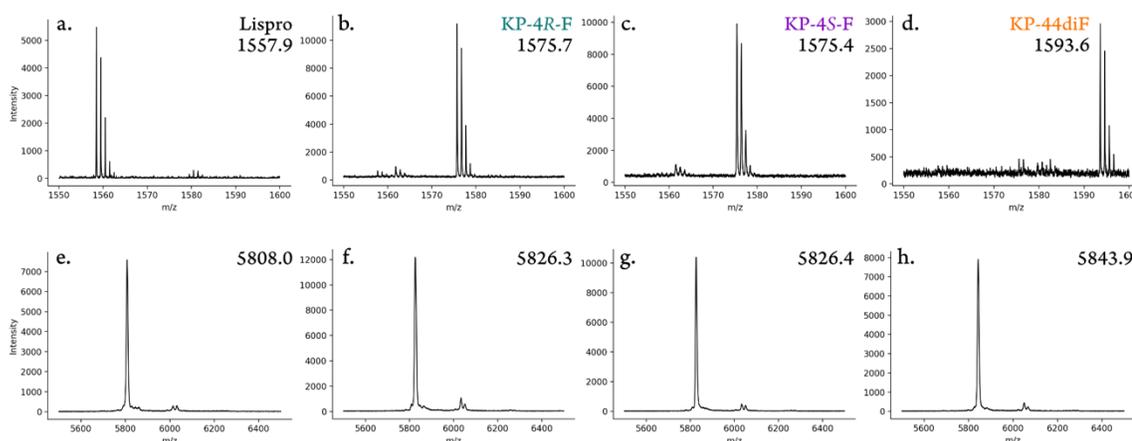


Figure 3.2. Mass spectrometry characterization of lispro variants. a-d. Characterization of proline analog incorporation. The solubilized inclusion body (containing proinsulin) after expression in the presence of proline (a), 4R-F (b), 4S-F (c), 44-diF (d) was digested with Glu-C and analyzed by MALDI-TOF MS. The peptide that contains position B29 of mature lispro is $^{50}\text{RGFFYTK}\underline{\text{P}}\text{TRRE}$ (expected $m/z = 1557.8$). e-h. MALDI-TOF characterization of mature and purified insulin variants: insulin lispro (e), KP-4R-F (f), KP-4S-F (g), and KP-44diF (h). The larger molecular weight peaks (~ 6050 Da) present in these spectra correspond to adducts of the sinapic acid matrix.

To verify the ability of each ncPro-containing insulin lispro to reduce blood glucose *in vivo*, insulins were injected subcutaneously into diabetic mice, and blood glucose was monitored over the course of 2.5 h. Mouse models allow determination of activity but not time to onset of action.²⁷ Because the B-chain C-terminus does not participate in binding to the insulin receptor,²⁸ we did not expect substitution at position B29 to affect bioactivity. Indeed, all lispro variants retained activity (Figure 3.3a).

The secondary structure of each insulin lispro variant was probed by circular dichroism spectroscopy (Figure 3.3b-e). Compared to human insulin, which exists as a dimer at $60 \mu\text{M}$, the monomeric insulin lispro exhibits increased negative ellipticity at 208 nm, and

decreased negative ellipticity at 222 nm, consistent with previous reports.⁸ The CD spectra of all fluorinated lispro variants were similar to that of lispro, suggesting that fluorination does not alter the secondary structure of lispro. Analytical ultracentrifugation confirmed this preference for the monomer state across all lispro variants (Figure 3.3f, Table 3.S3).

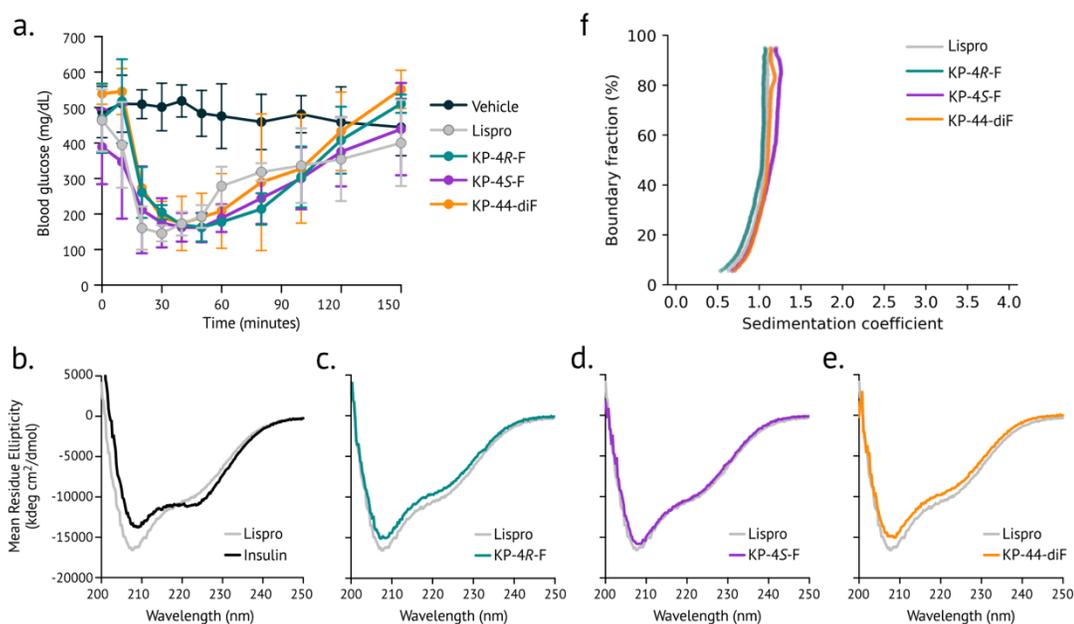


Figure 3.3. Bioactivity, circular dichroism spectroscopy, and sedimentation of lispro variants.

a. Changes in blood glucose concentrations over time after diabetic mice were injected subcutaneously with lispro variants. **b-e.** Far-UV circular dichroism spectra (60 μ M lispro in 10 mM phosphate buffer, pH 8.0). The spectrum of lispro is overlaid with that of human insulin (**b.**), KP-4R-F (**c.**), KP-4S-F (**d.**), and KP-44diF. **f.** Sedimentation velocity of lispro variants (125 μ M) in 100 mM phosphate, pH 8.0.

We measured the half-life ($t_{1/2}$) for dissociation of the hexamer form of each insulin variant by monitoring the decrease in negative ellipticity at 222 nm over time after dilution (Figure 3.4a). We formulated insulin and lispro variants under conditions that mimic the pharmaceutical formulation, using resorcinol as the phenolic ligand to slow dissociation.²⁹

Under these conditions (600 μM insulin, 25 mM resorcinol, 250 μM ZnCl_2), insulin lispro dissociated significantly more rapidly than human insulin ($t_{1/2} = 11.3 \pm 3.8$ and 30.2 ± 2.8 s, respectively), as anticipated. Accurately measuring dissociation behavior faster than lispro is expected to be challenging due to the resolution of this experiment, though comparisons to human insulin can still be made. We found that all fluorinated lispro variants dissociated similarly to proline-containing lispro (Figure 3.4b-g, Table 3.S3).

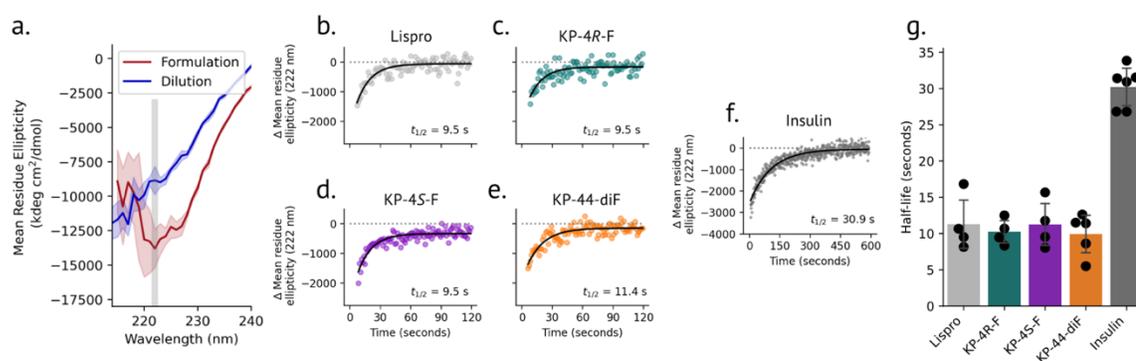


Figure 3.4. Hexamer dissociation kinetics of lispro variants. **a.** Equilibrium CD spectra of lispro before and after dilution. To measure dissociation kinetics, the decrease in negative ellipticity at 222 nm was monitored over time after dilution. **b-f.** Representative dissociation kinetics measurements for lispro (b), KP-4R-F (c), KP-4S-F (d), KP-44diF (e), and insulin (f). Note the extended x-axis for insulin in panel f. **g.** Summary of dissociation half-life values.

Each insulin lispro variant was subjected to continuous, vigorous shaking at 37°C to assess its stability against physical denaturation; fibrillation was monitored with the dye thioflavin T (ThT), (Figure 3.5a). Under these conditions, the fibrillation lag time of insulin lispro (10.9 ± 2.2 h) was shortened compared to that of human insulin (16.7 ± 4.1 h; see Chapter II), consistent with the notion that fibrillation proceeds from the monomer state.³⁰ Replacing ProB29 with 4R-F (10.3 ± 2.4 h) or 44-diF (9.0 ± 1.3 h) did not significantly change the fibrillation lag time. Notably, KP-4S-F was stabilized against fibril formation

(17.9 ± 0.8 h). We note that 4S-F favors the *endo* ring pucker (proline and the other ncPro variants are predominantly *exo*) and more strongly prefers the *cis* isomer than the other proline analogs examined in this study.^{14,31} Perhaps these conformational preferences at the C-terminus of the B-chain inhibit the formation of nuclei that precede fibril formation. TEM analysis of the lispro fibrils revealed similar morphology among samples (Figure 3.5b-e).

We probed the disorder of each lispro variant using the dye 8-anilino-1-naphthalenesulfonic acid (ANS), which exhibits a blue-shift in its emission maximum upon binding to hydrophobic patches of proteins indicative of a molten globule state.³² The emission spectra of all lispro variants were similar upon addition of ANS (Figure 3.5f), suggesting that the observed differences in fibrillation propensity were not due to variations in protein disorder.

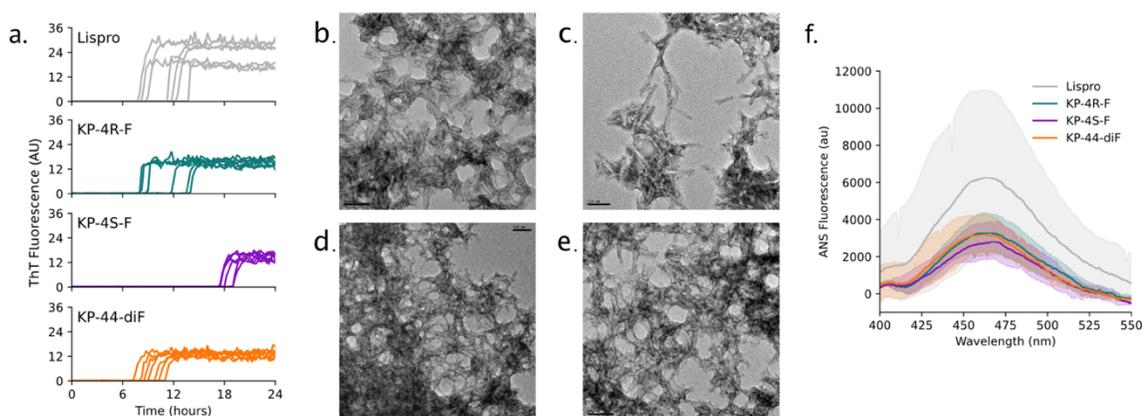


Figure 3.5. Fibrillation of lispro variants. a. Lispro variants (60 μ M in 100 mM phosphate buffer, pH 8.0) were incubated at 37°C with vigorous shaking; fibril formation was monitored by ThT fluorescence. b-e. TEM images of lispro (b), KP-4R-F (c), KP-4S-F (d), and KP-44-diF (e) aggregates. f. ANS emission spectra of lispro variants (1 μ M lispro variant labeled with 5 μ M ANS in 100 mM phosphate buffer, pH 8.0).

In conclusion, we report the synthesis of insulin lispro variants fluorinated at ProB29. While most properties of lispro variants remain unchanged, we find that 4*S*-fluorination of ProB29 in insulin lispro stabilizes the protein against physical denaturation. Approaches to protect insulin against denaturation often rely upon promoting oligomerization.^{24,33} However, here we observe stabilization of a monomeric FAI without changes in association behavior. Similar stabilization from the monomer state was observed for an insulin lispro variant that contains a 3-iodo substituent at TyrB26.³⁴ In both cases, the mechanism for stabilization is not fully established. Protein behavior is complex: ncPro substitution can impart long-range conformational effects,²³ and the precise molecular mechanism of insulin fibrillation is not fully understood.³⁰

Regardless of the mechanism, the discovery of a monomeric insulin with an extended lag time to fibril formation (KP-4*S*-F) is relevant to the design of improved insulin therapies. Current insulin formulations must be stored at 4°C and have limited shelf-lives. The inherent sensitivity of insulin to chemical and physical denaturation is especially problematic for long-term storage in continuous subcutaneous insulin infusion (CSII) pumps,³⁵ and in efforts to provide insulin-replacement therapies to individuals without regular access to refrigeration.³⁶

The results of this work highlight the ability of non-canonical amino acid mutagenesis to modulate pharmaceutically-relevant properties of therapeutic proteins. ncPro mutagenesis allows for modification of insulin lispro at position B29 without losing the unique conformational restrictions characteristic of proline, and the fluorinated insulin lispro

variants introduced here can be produced through straightforward bacterial expression. Future work is needed to more thoroughly elucidate the molecular origins for the stabilization of KP-4S-F.

3.4 Materials and methods

3.4.1 Chemicals

All chemicals were purchased from MilliporeSigma unless otherwise indicated. 4R-fluoroproline (4R-F) and 4S-fluoroproline (4S-F) were purchased from Bachem. 4,4-difluoroproline (44-diF) was purchased from Synthonix. All proline analogs were used as received.

3.4.2 Enzymes

Gibson Assembly enzymes were purchased as the Repliqa HiFi assembly mix from Quantabio. Trypsin was purchased from MilliporeSigma. Carboxypeptidase B was purchased from Worthington Biochemical. Glu-C peptidase was purchased from Promega.

3.4.3 Strains and plasmids

The proline-auxotrophic *E. coli* strain CAG18515 was obtained from the Coli Genetic Stock Center (CGSC) at Yale University. Strain DH10B was used for all cloning applications.

The plasmid pQE80_H27R-PI-KP_proS contains an IPTG-inducible proinsulin-lispro gene and an *E. coli* prolyl-tRNA synthetase gene controlled by its endogenous promoter. Proinsulin is translationally fused to an N-terminal leader peptide (H27R) that increases

expression yields,³⁷ and a 10x-his tag to facilitate proinsulin enrichment after refolding. The ProB28-LysB29 inversion present in insulin lispro was installed in plasmid pQE80_H27R-PI_proS (see Chapter II). We used a two-part Gibson Assembly approach with the two sets of primers described below. The two overlap regions were at the site of mutation, and within the ampicillin resistance gene selection marker. Correct installation of the desired mutation was verified by Sanger sequencing.

3.4.4 Primers

DNA oligos were purchased from Integrated DNA Technologies (IDT). Nucleotides responsible for the proline-lysine inversion are underlined.

Gibson assembly fragment 1:

SLB3099_GA-fwd: CGTGGTTTCTTCTACACGAAACCGACCCGCCGTGAAG

AmpR-GA_rev: GACAGTAAGAGAATTATGCAGTG

Gibson assembly fragment 2:

AmpR-GA_fwd: GCAGTGTTATCACTCATGG

SLB3099_GA_rev: CAGCTTACGGCGGGTCGGTTTCGTGTAGAAGAAACCACG

3.4.5 Nucleotide and amino acid sequences

H27R-PI-KP: The mutations responsible for the proline-lysine inversion are underlined.

ATGACAATGATCACTAATTCACCCGAGATTTCCCACCATCATCATCATCA
CCACCACCATCAGTTGATCTCGGAGGCCGTTTTGTGAACCAGCACCTGTGCG
GTAGCCACCTGGTGAAGCTCTGTACCTGGTTTGCGGTGAGCGTGGTTTCTTC
TACACGAAACCGACCCGCCGTGAAGCTGAAGATCTGCAGGTGGGCCAGGTA

GAACTGGGCGGTGGTCCGGGTGCCGGCTCTCTGCAACCGCTGGCACTGGAAG
 GTTCCCTGCAAGCGCGTGGTATCGTAGAGCAGTGCTGTACTTCTATCTGCTCC
 CTGTACCAGCTGGAGAACTACTGTAATTAA

The sequence of the H27R leader peptide is underlined, proinsulin is in **bold**. The A-chain and B-chain in mature insulin are colored **red** and **blue**, respectively.

MTMITNSPEISHHHHHHHHHHQLISEAR**FVNQHLCGSHLVEALYLVCGERGFF**
YTKPTRREAEDLQVGQVELGGPGAGSLQPLALEGSLQARGIVEQCCTSICS
LYQLENYCN

proS: the endogenous *proS* promoter is underlined, the coding sequence is in UPPERCASE.

attcaagccttctcttttgacatttctttgcactggtaaactaaatcactttttttgtcccaggctcgcttgagcctgttctaccttc
 aactggaaccgtaacaacATGCGTACTAGCCAATACCTGCTCTCCACTCTCAAGGAGAC
 ACCTGCCGACGCCGAGGTGATCAGCCATCAGCTGATGCTGCGCGCCGGGATG
 ATCCGCAAGCTGGCCTCCGGGTTATATACCTGGCTGCCGACCGGCGTGCGCG
 TTCTGAAAAAAGTCGAAAACATCGTGCGTGAAGAGATGAACAACGCCGGTGC
 GATCGAGGTGTCGATGCCGGTGGTTCAGCCAGCCGATTTGTGGCAAGAGAGT
 GGTCGTTGGGAACAGTACGGTCCGGAAGTCTGCTGCGTTTTGTTGACCGTGGCG
 AGCGTCCGTTTCGTACTIONCGGCCAACTCATGAAGAAGTTATCACTGACCTGATT
 CGTAACGAGCTTAGCTCTTACAAACAGCTGCCGCTGAACTTCTATCAGATCCA
 GACCAAGTTCCGCGACGAAGTGCCTCCGCGTTTCGGCGTCATGCGTTCCCGC
 GAATTCCTGATGAAAGATGCTTACTCTTTCCATACTTCTCAGGAATCCCTGCA
 GGAAACCTACGATGCAATGTATGCGGCCTACAGCAAAATCTTCAGCCGCATG
 GGGCTGGATTTCCGCGCCGTACAAGCCGACACCGGTTCTATCGGCGGCAGCG
 CCTCTCACGAATTCAGGTGCTGGCGCAGAGCGGTGAAGACGATGTGGTCTT
 CTCCGACACCTCTGACTATGCAGCGAACATTGAACTGGCAGAAGCTATCGCG
 CCGAAAGAACC GCGCGCTGCTGCTACCCAGGAAATGACGCTGGTTGATACGC
 CGAACGCGAAAACCATCGCGGAACTGGTTGAACAGTTCAATCTGCCGATTGA

GAAAACGGTTAAGACTCTGCTGGTTAAAGCGGTTGAAGGCAGCAGCTTCCCG
CAGGTTGCGCTGCTGGTGCGCGGTGATCACGAGCTGAACGAAGTTAAAGCAG
AAAAACTGCCGCAGGTTGCAAGCCCGCTGACTTTCGCGACCGAAGAAGAAAT
TCGTGCCGTGGTTAAAGCCGGTCCGGGTTCACTGGGTCCGGTAAACATGCCG
ATTCCGGTGGTGATTGACCGTACCGTTGCGGCGATGAGTGATTTTCGCTGCTGG
TGCTAACATCGATGGTAAACACTACTTCGGCATCAACTGGGATCGCGATGTC
GCTACCCCGGAAGTTGCAGATATCCGTAACGTGGTGGCTGGCGATCCAAGCC
CGGATGGCCAGGGTAGGCTGCTGATCAAACGTGGTATCGAAGTTGGTCACAT
CTCCAGCTGGGTACCAAGTACTCCGAAGCACTGAAAGCCTCCGTACAGGGT
GAAGATGGCCGTAACCAAATCCTGACGATGGGTTGCTACGGTATCGGGGTAA
CGCGTGTGGTAGCTGCGGCGATTGAGCAGAACTACGACGAACGAGGCATCGT
ATGGCCTGACGCTATCGCGCCGTTCCAGGTGGCGATTCTGCCGATGAACATG
CACAAATCCTTCCGCGTACAAGAGCTTGCTGAGAACTGTACAGCGAACTGC
GTGCACAAGGTATCGAAGTGCTGCTGGATGACCGCAAAGAGCGTCCGGGCGT
GATGTTTGCTGATATGGAAGTATCGGTATTCCGCACACTATTGTGCTGGGCG
ACCGTAACCTCGACAACGACGATATCGAATATAAATATCGTCGCAACGGCGA
GAAACAGTTAATTAAGACTGGTGACATCGTCGAATATCTGGTGAAACAGATT
AAAGGCTGA

MRTSQYLLSTLKETPADAEVISHQLMLRAGMIRKLASGLYTWLPTGVRVLKKVE
NIVREEMNAGAIIEVSMPPVQPADLWQESGRWEQYGPPELLRFVDRGERPFVLG
PTHEEVITDLIRNELSSYKQLPLNFYQIQTKFRDEVPRFGVMRSREFLMKDAYSF
HTSQESLQETYDAMYAAYSKIFSRMGLDFRAVQADTGSIGGSASHEFQVLAQSG
EDDVVFSDTSDYAANIELAEAIAPKEPRAAATQEMTLVDTPNAKTIAELVEQFNL
PIEKTVKTLLVKAVEGSSFPQVALLVRGDHELNEVKAELPQVASPLTFATEEEIR
AVVKAGPGSLGPVNMPIPVVIDRTVAAMSDFAAAGANIDGKHYFGINWDRDVAT
PEVADIRNVVAGDPSPDGQGRLLIKRGIEVGHIFQLGTKYSEALKASVQGEDGRN
QILTMGCYGIGVTRVVAIAEQNYDERGIVWPDAIAPFQVAILPMNMHKSFRVQ
ELAELYSELRAQGIEVLLDDRKERPGVMFADMELIGIPHTIVLGDRNLDNDIE
YKYRRNGEKQLIKTG DIVEYLVKQIKG

3.4.6 Proinsulin expression

A single colony of *E. coli* strain CAG18515 harboring plasmid pEQ80_H27R-PI-KP_proS was used to inoculate 70 mL of Luria Bertani (LB) medium containing ampicillin, and the culture was grown overnight at 37°C to stationary phase. The overnight culture was used to inoculate 5 L (as 4x1.25 L cultures) of 1x Andrew's Magical Medium (AMM),³⁸ a defined medium containing all 20 proteinogenic amino acids, in 2.8 L Fernbach flasks. The composition of AMM was the following: 3.60 g L⁻¹ glucose, 3.5 g L⁻¹ KH₂PO₄, 6.56 g L⁻¹ K₂HPO₄•3H₂O, 3.5 g L⁻¹ (NH₄)₂HPO₄, 8.37 g L⁻¹ MOPS, 0.72 g L⁻¹ tricine, 2.92 g L⁻¹ NaCl, 0.51 g L⁻¹ NH₄Cl, 0.26 g L⁻¹ MgCl₂•7H₂O, 50 mg L⁻¹ K₂SO₄, 0.246 mg L⁻¹ MgSO₄•7H₂O, 12.3 mg L⁻¹ CaCl₂•2H₂O, 2.8 mg L⁻¹ FeSO₄•7H₂O, 0.5 mg L⁻¹ thiamine, 24 µg L⁻¹ boric acid, 1 µg L⁻¹ trace metals (Cu²⁺, Mn²⁺, Zn²⁺, MoO₄²⁻), and 50 mg L⁻¹ each amino acid.

When growth reached mid-exponential phase (OD₆₀₀ ~0.8), the culture was subjected to a medium shift: cells were pelleted via centrifugation (5 kg, 5 min, 4°C) and washed twice with ice-cold 0.9% NaCl. Washed cells were resuspended in 1 L of 1.25x AMM –Pro, a 1.25x concentrated form of AMM that omits proline. Cells were incubated for 30 min at 37°C to deplete residual proline, after which 250 mL of a solution containing 2.5 mM ncPro and 1.5 M NaCl was added (0.5 mM ncPro and 0.3 M NaCl working concentrations). For the incorporation of 44-diF, the concentration of NaCl was 2.5 M (0.5 M working concentration). After 30 min of incubation at 37°C to allow for ncPro uptake, proinsulin expression was induced by the addition of 1 mM IPTG. Cultures were incubated overnight

at 37°C, after which cells were harvested via centrifugation and stored at -80°C until further processing.

Proline-containing proinsulins were expressed in rich medium: proinsulin-lispro was expressed using strain CAG18515 harboring plasmid pQE80-H27R-PI-KP_proS in 6 L (as 6 x 1.0 L cultures) of Terrific Broth (TB). 1 mM IPTG was added at mid-log phase ($OD_{600} \sim 0.8$) to induce proinsulin expression. Cultures were incubated at 37°C for 3 h, after which cells were harvested via centrifugation and stored at -80°C until further processing.

3.4.7 Proinsulin refolding

Cell pellets were warmed from -80°C to room temperature and resuspended in 5 mL IB buffer (50 mM tris, 100 mM NaCl, 1 mM EDTA, pH 8.0) per gram cell pellet. 1 mg L⁻¹ lysozyme and 1 mM PMSF were added, and the slurry sat on ice for 30 min before cells were lysed via sonication. The lysate was centrifuged (14 kg, 30 min, 4°C) and the soluble fraction was discarded. The pellet was washed twice with IB buffer + 1% Triton X-100, once with IB buffer, and once with water; this final step required centrifugation for 45 min. The washed inclusion body pellet was resuspended in a minimal amount of water, and the mass of proinsulin in the inclusion body pellet was estimated by SDS-PAGE.

In preparation for proinsulin refolding, the inclusion body was resuspended in 3 M urea and 10 mM cysteine in water, such that the proinsulin concentration was 1 mg proinsulin per L total slurry. To dissolve proinsulin, the pH was adjusted to 12 and sample stirred for 1 h at room temperature. At this stage, ncPro incorporation was assessed by MALDI-TOF, which is described in section 3.4.10 below. The solubilized proinsulin solution was diluted

ten-fold into refold buffer (10 mM CAPS, pH 10.6) that had been pre-cooled to 4°C. The pH of the refold solution was adjusted to 10.7 and the sample stored at 4°C; care was taken to ensure that the solution pH remained between 10.6 and 10.8 throughout the refolding process. Proinsulin refolding progress was monitored by reverse-phase HPLC, and usually reached completion within 50 h.

Proinsulin was enriched from the refold solution after adjusting the pH to 8.0 and incubating the sample overnight with Ni-NTA resin and 10 mM imidazole. The resin was washed with wash buffer (25 mM imidazole in PBS, pH 8.0), and proinsulin was eluted with elution buffer (250 mM imidazole in PBS, pH 8.0). Fractions containing proinsulin were combined and extensively dialyzed against 10 mM sodium phosphate, pH 8.0.

3.4.8 Lispro maturation and purification

Refolded and dialyzed proinsulin was warmed to 37°C and digested with trypsin (20 U mL⁻¹) and carboxypeptidase-B (10 U mL⁻¹) at 37°C for 2.5 h to remove the N-terminal tag and C chain. Digestion was halted by adjusting the pH to ~3 with 6 N HCl.

Lispro variants were immediately purified after proteolysis by reverse-phase HPLC on a C₄ column (Penomenex Jupiter 5 µm particle size, 300 Å pore size, 250x10 mm) using 0.1% TFA in water (solvent A) and 0.1% TFA in acetonitrile (solvent B) as mobile phases. A gradient of 25-32% solvent B was applied over 65 min, and fractions containing lispro were collected. Samples for purity analysis were removed at this stage; the remaining portion of the fraction was lyophilized. Each lispro fraction was analyzed by analytical reverse-phase HPLC, MALDI-TOF MS (Figure 3.2e-h), and SDS-PAGE (Figure 3.S1) to

verify sample quality and ensure $\geq 95\%$ purity for all downstream analyses. Lyophilized powders were stored at -20°C until further use.

3.4.9 Preparation of lispro variants for mouse experiments

We used an earlier version of this protocol to express lispro variants used in mouse studies.^{24,25} Briefly, the hexahistidine-tagged version of proinsulin (lacking the H27R leader peptide) was expressed in M9 medium using the strain CAG18515 harboring plasmid pQE80PI-KP_proS. The medium shift described above was used to incorporate the ncPro residue of interest, and proinsulin was expressed for 2.5 h. After cell lysis, proinsulin was solubilized with 8 M urea and isolated from the washed inclusion body fraction by Ni-NTA under denaturing conditions. Refolding was achieved after oxidative sulfiteolysis by gentle agitation at 12°C overnight in 0.5 mM BME, 0.3 M urea, 50 mM glycine, pH 12. After dialysis, lyophilization, and proteolysis by trypsin and carboxypeptidase B, mature lispro variants were purified by reverse-phase HPLC.

During the course of this work, we found that the refolding protocol described in section 3.4.7 resulted in an approximately seven-fold increase in refolding yield, and higher purity samples.

3.4.10 MALDI-TOF MS

To assess levels of incorporation of ncPros into the corresponding proinsulins, samples were subjected to Glu-C digestion, which results in a peptide fragment containing ProB29 (⁵⁰RGFFYTKPTRRE). A 20 μL aliquot of the solubilized proinsulin-containing inclusion body fraction was subjected to cysteine reduction (5 mM DTT, 55°C for 20 min) and

alkylation (15 mM iodoacetimide, RT for 15 min in the dark), prior to 10-fold dilution into 100 mM NH_4HCO_3 , pH 8.0 (100 μL final volume). Digestion was started with addition of 0.6 μL Glu-C ($0.5 \mu\text{g} \mu\text{L}^{-1}$ in ddH_2O) at 37°C for 2.5 h. The digestion reaction was quenched by adding 10 μL of 5% TFA. Peptides were desalted using ZipTip C_{18} columns (MilliporeSigma) according to the manufacturer's protocol. Desalted peptides (in 50% acetonitrile, 0.1% TFA) were diluted 3:1 into the matrix solution (α -cyanohydroxycinnamic acid in 50% ACN, 0.1% TFA) and analyzed by MALDI-TOF MS. Analog incorporation was calculated by comparing the area under the curve (AUC) of the ncPro form of the peptide ($m/z = 1576$ for 4*R*-F and 4*S*-F, and 1595 for 44-diF) with the AUC of the canonical proline peptide ($m/z = 1558$).

HPLC-purified insulin and lispros were analyzed as full-length, mature proteins. Aliquots directly from HPLC purification (~30% ACN, 0.1% TFA) were mixed 1:1 with matrix solution (sinapic acid in 30% ACN, 0.1% TFA) before analysis by MALDI-TOF MS.

3.4.11 Reduction of blood glucose in diabetic mice

NODscid mice (NOD.CB17-Prkdcscid/J) were obtained from Jax Mice (Bar Harbor, ME). Mice were maintained under specific pathogen-free conditions, and experiments were conducted according to procedures approved by the Institutional Animal Care and Use Committee (IACUC) at the City of Hope. Adult (8-12 week old) male NODscid mice were injected intraperitoneally ($45 \text{ mg kg}^{-1} \text{ day}^{-1}$ for 3 days) with freshly prepared streptozotocin (STZ) in 50 mM citrate buffer, pH 4.5 to induce diabetes. Diabetes was confirmed 3 weeks after the last dose of STZ by detection high glucose levels ($200\text{-}600 \text{ mg dL}^{-1}$) as measured

by a glucomonitor (Freestyle, Abbott Diabetes Care, Alameda, CA) in blood sampled from the lateral tail vein. Lispro analogs were diluted to $100 \mu\text{g mL}^{-1}$ in formulation buffer (1.6 mg mL^{-1} *m*-cresol, 0.65 mg mL^{-1} phenol, 3.8 mg mL^{-1} sodium phosphate pH 7.4, 16 mg mL^{-1} glycerol, $0.8 \mu\text{g mL}^{-1}$ ZnCl_2). Lispro analogs were injected ($35 \mu\text{g kg}^{-1}$) subcutaneously at the scruff and blood glucose was measured at 0, 10, 20, 30, 40, 50, 60, 80, 100, 120, and 150 min.

3.4.12 Circular dichroism spectroscopy

Equilibrium measurements: The circular dichroism spectra of lispro samples ($60 \mu\text{M}$ in 100 mM sodium phosphate, pH 8.0) were measured at 25°C in 1 mm quartz cuvettes on an Aviv Model 430 Circular Dichroism Spectrophotometer using a step size of 0.5 nm and averaging time of 1 s . A reference buffer spectrum was subtracted from each sample spectrum.

Kinetic measurements: Insulin samples in 100 mM sodium phosphate buffer pH 8.0 were dialyzed overnight against 28.6 mM tris buffer, pH 8.0 (Slide-A-Lyzer dialysis cassettes, 3.5 kDa MWCO, ThermoFisher). Insulins were formulated as the following: $600 \mu\text{M}$ insulin, $250 \mu\text{M}$ ZnCl_2 , 25 mM resorcinol, 25 mM tris buffer, pH 8. To a stirred buffer solution containing 2.98 mL of 25 mM tris, pH 8.0 in a 10 mm quartz cuvette was injected $20 \mu\text{L}$ of the insulin formulation (150 -fold dilution). Ellipticity was monitored at 222 nm over 120 s (1 s kinetic interval, 0.5 s time constant) at 25°C . A typical run led to a rapid drop in CD signal as mixing occurred ($\sim 5 \text{ s}$), then a gradual rise to an equilibrium ellipticity representative of an insulin monomer. Data preceding the timepoint with the greatest

negative ellipticity (representing the mixing time) were omitted from further analysis. Runs were discarded if the maximum change in mean residue ellipticity from equilibrium did not exceed $750 \text{ deg cm}^2 \text{ dmol}^{-1}$, which was indicative of poor mixing. The remaining data were fit to a mono-exponential function using Scipy (Python); data presented here are from at least two separate HPLC fractions, measured on two different days.

An equilibrium spectrum for each protein was obtained after dilution; all spectra were indicative of an insulin monomer. The CD spectrum of lispro under pre-dilution formulation conditions was obtained using a 0.1 mm quartz cuvette. In each case, a blank spectrum containing all buffers and ligands was subtracted from the sample spectrum.

3.4.13 Fibrillation

Lispro samples ($60 \mu\text{M}$ in 100 mM sodium phosphate, pH 8.0) were centrifuged at $22,000 \text{ g}$ for 1 h at 4°C , prior to the addition of $1 \mu\text{M}$ thioflavin T (ThT). Each lispro ($200 \mu\text{L}$) was added to a 96-well, black, clear bottom plate (Greiner Bio-One) and sealed. Samples were shaken continuously at 960 rpm on a Varioskan multimode plate reader at 37°C , and fluorescence readings were recorded every 15 min (444 nm excitation, 485 nm emission). Fibrillation runs were performed on at least two separate HPLC fractions, each in triplicate or quadruplicate, and on two different days. The growth phase of each fibrillation replicate was fit to a linear function, and fibrillation lag times were reported as the x-intercept of this fit. Fibril samples were stored at 4°C until analysis by TEM.

3.4.14 Transmission electron microscopy

Lispro fibrils were centrifuged (5 kg, 1 min), then washed twice and resuspended in ddH₂O. Fibrils were stained with 2% uranyl acetate on a 300-mesh formvar/carbon coated copper grid (Electron Microscopy Sciences) and imaged on a Tecnai T12 LaB6 120 eV transmission electron microscope.

3.4.15 ANS fluorescence

1 μ M lispro variant was mixed with 5 μ M ANS in 100 mM phosphate buffer, pH 8.0. Fluorescence emission spectra were measured in 1 cm quartz cuvettes at ambient temperature using a PTI QuantaMaster fluorescence spectrofluorometer. A 350 nm excitation wavelength and scan rate of 2 nm s⁻¹ were used. Measurements for each variant were performed in triplicate from three separate HPLC fractions.

3.4.16 Analytical ultracentrifugation

Lispro variants were formulated at 125 μ M in 100 mM phosphate buffer, pH 8.0. Velocity sedimentation experiments were performed at the Canadian Center for Hydrodynamics by Maduni Ranashinghe, with input from Borries Demeler. Data were analyzed with UltraScan III version 4.0 release 6606.³⁹ Sedimentation coefficients were determined from the enhanced van Holde-Weischet⁴⁰ analysis to generate diffusion-corrected sedimentation coefficient distributions.

3.5 Supplementary figures and tables

Figure 3.S1. Purity of lispro samples assessed by SDS-PAGE. HPLC-purified lispro (a), KP-4R-F (b), KP-4S-F (c), and KP-44diF were analyzed by SDS-PAGE to validate purity; shown are representative lanes corresponding to individual HPLC fractions.

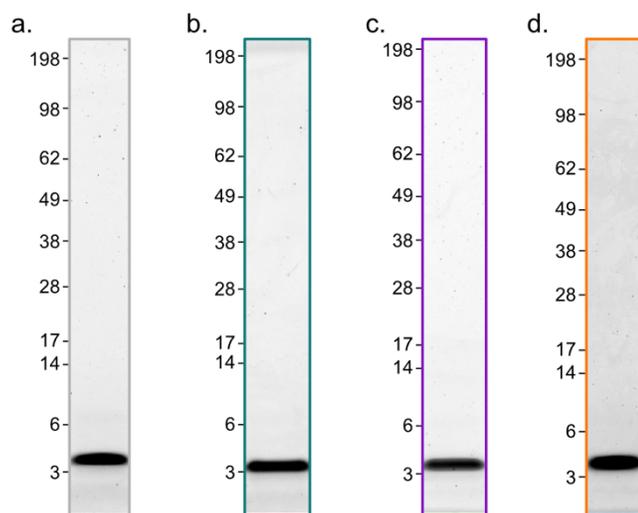


Figure 3.S2. Circular dichroism controls. At 60 μM , human insulin is expected to exist as a dimer at pH 8, monomer in 20% ethanol, and denatured in the presence of 8 M guanidinium chloride. These spectra are overlaid with equilibrium spectra collected before and after lispro dilution for kinetic CD measurements. Spectra below 210-215 nm were omitted for some samples due to high levels of buffer absorbance at these wavelengths.

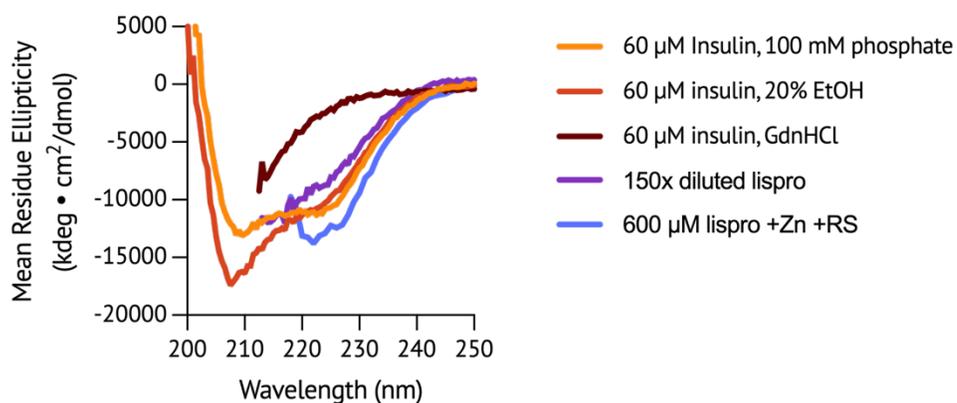


Table 3.S1. Mass spectrometry characterization of lispro variants

Protein	Digested peptide			Mature insulin	
	Expected m/z	Observed m/z	Incorporation efficiency	Expected m/z	Observed m/z
Lispro	1557.78	1558.0	–	5808.6	5807.7 ± 0.4
KP-4R-F	1575.82	1575.737 ± 0.008	0.961 ± 0.005	5826.6	5826.31 ± 0.03
KP-4S-F	1575.82	1575.53 ± 0.01	0.948 ± 0.033	5826.6	5826.5 ± 0.1
KP-44-diF	1593.81	1593.57 ± 0.01	0.912 ± 0.016	5844.6	5844.1 ± 0.7

Table 3.S2. Lispro yields

Protein	Proinsulin yield (mg L ⁻¹) [‡]	Approx. insulin yield (mg L ⁻¹)
Lispro [*]	33	n.d. [#]
KP-4R-F	32	3.4
KP-4S-F	7	2.2
KP-44-diF	8	2.4

[‡]Yields determined by measuring absorbance (280 nm) after Ni-NTA enrichment following proinsulin refolding.

^{*}Expressed in terrific broth (TB)

[#]n.d., not determined

Table 3.S3. Summary of lispro variant characterization

Protein	Ellipticity ratio (208/222 nm) [‡]	Hexamer dissociation t _{1/2} (s)	Fibrillation lag time (h) [‡]	ANS emission maximum (nm) [%]	Sedimentation coefficient (S) ^{&}
Lispro	1.63 ± 0.07	11.3 ± 3.8	10.9 ± 2.2	466 ± 2	1.0
KP-4R-F	1.62 ± 0.06	10.2 ± 1.6	10.3 ± 2.4	464 ± 4	1.0
KP-4S-F	1.59 ± 0.06	11.2 ± 3.3	17.9 ± 0.8	463 ± 6	1.1
KP-44-diF	1.61 ± 0.05	9.9 ± 2.9	9.0 ± 1.3	462 ± 5	1.1
Insulin	1.24 ± 0.03 [*]	30.2 ± 2.8	16.6 ± 4.1 [*]	470 ± 5 [*]	n.d. [#]

[‡]60 μM insulin, 100 mM phosphate buffer, pH 8.0

[%]1 μM insulin, 5 μM ANS, 100 mM phosphate buffer, pH 8.0

[&]125 μM insulin, 100 mM phosphate buffer, pH 8.0

^{*}Chapter II

[#]n.d., not determined

3.6 References

- (1) Haeusler, R. A.; McGraw, T. E.; Accili, D. Biochemical and Cellular Properties of Insulin Receptor Signalling. *Nat Rev Mol Cell Biol* **2018**, *19*, 31–44. <https://doi.org/10.1038/nrm.2017.89>.
- (2) World Health Organization. *Global Report on Diabetes*; 2016. <https://www.who.int/publications/i/item/9789241565257> (accessed 2023-04-02).
- (3) Goeddel, D. V.; Kleid, D. G.; Bolivar, F.; Heyneker, H. L.; Yansura, D. G.; Crea, R.; Hirose, T.; Kraszewski, A.; Itakura, K.; Riggs, A. D. Expression in *Escherichia coli* of Chemically Synthesized Genes for Human Insulin. *Proc Natl Acad Sci* **1979**, *76* (1), 106–110. <https://doi.org/10.1073/pnas.76.1.106>.
- (4) Derewenda, U.; Derewenda, Z.; Dodson, E. J.; Reynolds, C. D.; Smith, G. D.; Sparks, C.; Swenson, D. Phenol Stabilizes More Helix in a New Symmetrical Zinc Insulin Hexamer. *Nature* **1989**, *338*, 594–596. <https://doi.org/10.1038/338594a0>.
- (5) Birnbaum, D. T.; Kilcomons, M. A.; DeFelippis, M. R.; Beals, J. M. Assembly and Dissociation of Human Insulin and LysB28-ProB29-Insulin Hexamers: A Comparison Study. *Pharm Res* **1997**, *14* (1), 25–36. <https://doi.org/10.1023/a:1012095115151>.
- (6) Holleman, F.; Hoekstra, J. B. L. Insulin Lispro. *N Engl J Med* **1997**, *337* (3), 176–183. <https://doi.org/10.1056/NEJM199707173370307>.
- (7) Bakaysa, D. L.; Radziuk, J.; Havel, H. A.; Brader, M. L.; Li, S.; Dodd, S. W.; Beals, J. M.; Pekar, A. H.; Brems, D. N. Physicochemical Basis for the Rapid Time-Action of LysB28ProB29-Insulin: Dissociation of a Protein-Ligand Complex. *Protein Sci* **1996**, *5*, 2521–2531. <https://doi.org/10.1002/pro.5560051215>.
- (8) Brems, D. N.; Alter, L. A.; Beckage, M. J.; Chance, R. E.; Dimarchi, R. D.; Green, L. K.; Long, H. B.; Pekar, A. H.; Shields, J. E.; Frank, B. H. Altering the Association Properties of Insulin by Amino Acid Replacement. *Protein Eng* **1992**, *5* (6), 527–533. <https://doi.org/10.1093/protein/5.6.527>.
- (9) Brange, J.; Ribel, U.; Hansen, J. F.; Dodson, G.; Hanse, M. T.; Havelund, S.; Melberg, S. G.; Norris, F.; Norris, K.; Snel, L.; Sørensen, A. R.; Voigt, H. O. Monomeric Insulins Obtained by Protein Engineering and Their Medical Implications. *Nature* **1988**, *333*, 679–682. <https://doi.org/10.1038/333679a0>.
- (10) Rakatzi, I.; Ramrath, S.; Ledwig, D.; Dransfeld, O.; Bartels, T.; Seipke, G.; Eckel, J. A Novel Insulin Analog With Unique Properties: LysB3 ,GluB29 Insulin Induces Prominent Activation of Insulin Receptor Substrate 2, but Marginal Phosphorylation of Insulin Receptor Substrate 1. *Diabetes* **2003**, *52*, 2227–2238. <https://doi.org/10.2337/diabetes.52.9.2227>.
- (11) Mathieu, C.; Gillard, P.; Benhalima, K. Insulin Analogues in Type 1 Diabetes Mellitus: Getting Better All the Time. *Nat Rev Endocrinol* **2017**, *13*, 385–399. <https://doi.org/10.1038/nrendo.2017.39>.
- (12) Brems, D. N.; Brown, P. L.; Bryant, C.; Chance, R. E.; Green, L. K.; Long, H. B.; Miller, A. A.; Millican, R.; Shields, J. E.; Frank, B. H. Improved Insulin Stability through

- Amino Acid Substitution. *Protein Eng* **1992**, *5* (6), 519–525.
<https://doi.org/10.1093/protein/5.6.519>.
- (13) Lepore, M.; Pampanelli, S.; Fanelli, C.; Porcellati, F.; Bartocci, L.; Di Vincenzo, A.; Cordoni, C.; Costa, E.; Brunetti, P.; Bolli, G. B. Pharmacokinetics and Pharmacodynamics of Subcutaneous Injection of Long-Acting Human Insulin Analog Glargine, NPH Insulin, and Ultralente Human Insulin and Continuous Subcutaneous Infusion of Insulin Lispro. *Diabetes* **2000**, *49*, 2142–2148.
<https://doi.org/10.2337/diabetes.49.12.2142>.
- (14) Ganguly, H. K.; Basu, G. Conformational Landscape of Substituted Prolines. *Biophys Rev* **2020**, *12*, 25–39. <https://doi.org/10.1007/s12551-020-00621-8>.
- (15) Chin, J. W. Expanding and Reprogramming the Genetic Code. *Nature* **2017**, *550*, 53–60.
<https://doi.org/10.1038/nature24031>.
- (16) Renner, C.; Alefelder, S.; Bae, J. H.; Budisa, N.; Huber, R.; Moroder, L. Fluoroproline as Tools for Protein Design and Engineering. *Angew Chem Int Ed* **2001**, *40* (5), 923–925.
[https://doi.org/10.1002/1521-3773\(20010302\)40:5<923::AID-ANIE923>3.0.CO;2-#](https://doi.org/10.1002/1521-3773(20010302)40:5<923::AID-ANIE923>3.0.CO;2-#).
- (17) Holmgren, S. K.; Taylor, K. M.; Bretscher, L. E.; Raines, R. T. Code for Collagen's Stability Deciphered. *Nature* **1998**, *392*, 666–667. <https://doi.org/10.1038/33573>.
- (18) Lummis, S. C. R.; Beene, D. L.; Lee, L. W.; Lester, H. A.; Broadhurst, R. W.; Dougherty, D. A. *Cis-Trans* Isomerization at a Proline Opens the Pore of a Neurotransmitter-Gated Ion Channel. *Nature* **2005**, *438*, 248–252.
<https://doi.org/10.1038/nature04130>.
- (19) Crespo, M. D.; Rubini, M. Rational Design of Protein Stability: Effect of (2*S*,4*R*)-4-Fluoroproline on the Stability and Folding Pathway of Ubiquitin. *PLoS One* **2011**, *6* (5), e19425. <https://doi.org/10.1371/journal.pone.0019425>.
- (20) Naduthambi, D.; Zondlo, N. J. Stereoelectronic Tuning of the Structure and Stability of the Trp Cage Miniprotein. *J Am Chem Soc* **2006**, *128*, 12430–12431.
<https://doi.org/10.1021/ja0648458>.
- (21) Rubini, M.; Schärer, M. A.; Capitani, G.; Glockshuber, R. (4*R*)- and (4*S*)-Fluoroproline in the Conserved *Cis*-Prolyl Peptide Bond of the Thioredoxin Fold: Tertiary Structure Context Dictates Ring Puckering. *ChemBioChem* **2013**, *14* (9), 1053–1057.
<https://doi.org/10.1002/cbic.201300178>.
- (22) Torbeev, V. Y.; Hilvert, D. Both the *Cis-Trans* Equilibrium and Isomerization Dynamics of a Single Proline Amide Modulate β 2-Microglobulin Amyloid Assembly. *Proc Natl Acad Sci* **2013**, *110* (50), 20051–20056. <https://doi.org/10.1073/pnas.1310414110>.
- (23) Torbeev, V.; Ebert, M. O.; Dolenc, J.; Hilvert, D. Substitution of Proline₃₂ by α -Methylproline Preorganizes β 2-Microglobulin for Oligomerization but Not for Aggregation into Amyloids. *J Am Chem Soc* **2015**, *137* (7), 2524–2535.
<https://doi.org/10.1021/ja510109p>.

- (24) Fang, K. Y.; Lieblich, S. A.; Tirrell, D. A. Replacement of ProB28 by Pipecolic Acid Protects Insulin against Fibrillation and Slows Hexamer Dissociation. *J Polym Sci A Polym Chem* **2019**, *57* (3), 264–267. <https://doi.org/10.1002/pola.29225>.
- (25) Lieblich, S. A.; Fang, K. Y.; Cahn, J. K. B.; Rawson, J.; LeBon, J.; Teresa Ku, H.; Tirrell, D. A. 4*S*-Hydroxylation of Insulin at ProB28 Accelerates Hexamer Dissociation and Delays Fibrillation. *J Am Chem Soc* **2017**, *139*, 8384–8387. <https://doi.org/10.1021/jacs.7b00794>.
- (26) Breunig, S. L.; Tirrell, D. A. Incorporation of Proline Analogs into Recombinant Proteins Expressed in *Escherichia coli*. *Methods Enzymol* **2021**, *656*, 545–571. <https://doi.org/10.1016/BS.MIE.2021.05.008>.
- (27) Plum, A.; Agersø, H.; Andersen, L. Pharmacokinetics of the Rapid-Acting Insulin Analog, Insulin Aspart, in Rats, Dogs, Pigs, and Pharmacodynamics of Insulin Aspart in Pigs. *Drug Metab Dispos* **2000**, *28*, 155–160.
- (28) Menting, J. G.; Whittaker, J.; Margets, M. B.; Whittaker, L. J.; Kong, G. K.-W.; Smith, B. J.; Watson, C. J.; Žáková, L.; Kletvíková, E.; Jiracek, J.; Chan, S. J.; Steiner, D. F.; Dodson, G. G.; Brzozowski, A. M.; Weiss, M. A.; Ward, C. W.; Lawrence, M. C. How Insulin Engages Its Primary Binding Site on the Insulin Receptor. *Nature* **2013**, *493*, 241–245. <https://doi.org/10.1038/nature11781>.
- (29) Rahuel-Clermont, S.; French, C. A.; Kaarsholm, N. C.; Dunn, M. F. Mechanisms of Stabilization of the Insulin Hexamer through Allosteric Ligand Interactions. *Biochemistry* **1997**, *36* (19), 5837–5845. <https://doi.org/10.1021/bi963038q>.
- (30) Akbarian, M.; Yousefi, R.; Farjadian, F.; Uversky, V. N. Insulin Fibrillation: Toward Strategies for Attenuating the Process. *ChemCommun* **2020**, *56* (77), 11354–11373. <https://doi.org/10.1039/d0cc05171c>.
- (31) Shoulders, M. D.; Raines, R. T. Collagen Structure and Stability. *Annu Rev Biochem* **2009**, *78*, 929–958. <https://doi.org/10.1146/annurev.biochem.77.032207.120833>.
- (32) Semisotnov, G. V.; Rodionova, N. A.; Razgulyaev, O. I.; Uversky, V. N.; Gripas', A. F.; Gilmanishin, R. I. Study of the “Molten Globule” Intermediate State in Protein Folding by a Hydrophobic Fluorescent Probe. *Biopolymers* **1991**, *31* (1), 119–128. <https://doi.org/10.1002/BIP.360310111>.
- (33) Brange, J.; Andersen, L.; Laursen, E. D.; Meyn, G.; Rasmussen, E. Toward Understanding Insulin Fibrillation. *J Pharm Sci* **1997**, *86* (5), 517–525. <https://doi.org/10.1021/js960297s>.
- (34) Pandyarajan, V.; Phillips, N. B.; Cox, G. P.; Yang, Y.; Whittaker, J.; Ismail-Beigi, F.; Weiss, M. A. Biophysical Optimization of a Therapeutic Protein by Nonstandard Mutagenesis: Studies of an Iodo-Insulin Derivative. *J Biol Chem* **2014**, *289* (34), 23367–23381. <https://doi.org/10.1074/jbc.M114.588277>.
- (35) Reichel, A.; Rietzsch, H.; Köhler, H.; Pfützner, A.; Gudat, U.; Schulze, J. Cessation of Insulin Infusion at Night-Time during CSII-Therapy: Comparison of Regular Human Insulin and Insulin Lispro. *Exp Clin Endocrinol Diabetes* **1998**, *106* (03), 168–172. <https://doi.org/10.1055/s-0029-1211971>.

- (36) Weiss, M. A. Design of Ultra-Stable Insulin Analogues for the Developing World. *J Health Specialties* **2013**, *1* (2), 59–70. <https://doi.org/10.4103/1658-600X.114683>.
- (37) Min, C. K.; Son, Y. J.; Kim, C. K.; Park, S. J.; Lee, J. W. Increased Expression, Folding and Enzyme Reaction Rate of Recombinant Human Insulin by Selecting Appropriate Leader Peptide. *J Biotechnol* **2011**, *151* (4), 350–356. <https://doi.org/10.1016/j.jbiotec.2010.12.023>.
- (38) He, W.; Fu, L.; Li, G.; Andrew Jones, J.; Linhardt, R. J.; Koffas, M. Production of Chondroitin in Metabolically Engineered *E. coli*. *Metab Eng* **2015**, *27*, 92–100. <https://doi.org/10.1016/J.YMBEN.2014.11.003>.
- (39) Demeler, B.; Gorbet, G. E. Analytical Ultracentrifugation Data Analysis with Ultrascan-III. *Analytical Ultracentrifugation: Instrumentation, Software, and Applications* **2016**, 119–143. https://doi.org/10.1007/978-4-431-55985-6_8.
- (40) Demeler, B.; Van Holde, K. E. Sedimentation Velocity Analysis of Highly Heterogeneous Systems. *Anal Biochem* **2004**, *335* (2), 279–288. <https://doi.org/10.1016/J.AB.2004.08.039>.

INCORPORATION OF PHOTO-PROLINE INTO PROTEINS IN *ESCHERICHIA COLI*

4.1 Contributions

Mona Shahgholi performed mass spectrometry of proinsulin peptides to determine photo-proline incorporation efficiencies. Jeff Jones, Brett Lomenick, Ting Yu Wang, and Tsui-Fen Chou assisted with determining photo-pro photoconversion by mass spectrometry. Edgar Manriquez-Sandoval and Anneliese Faustino performed PPIase enrichments, fractionation, and mass spectrometry, with input from Stephen D. Fried. Stephanie L. Breunig prepared all plasmids, performed all other experiments, participated in proteomics data analysis, and analyzed all other data. We thank Gracie Wang for assistance with UV irradiation.

4.2 Abstract

Photo-crosslinking non-canonical amino acids contain light-activated side chains that form covalent crosslinks to nearby molecules upon irradiation. Diverse non-canonical residues carrying photo-active moieties have characterized interactions between proteins and other biomolecules. Here, we report the ribosomal incorporation of photo-proline, a diazirine-containing proline analog, into recombinantly-expressed proinsulin in *Escherichia coli*. We found that up to 74% of proline residues could be replaced with photo-proline, and sought to leverage this photo-crosslinking amino acid to detect protein-protein interactions at proline positions in live cells. However, initial efforts did not detect crosslinking between

E. coli peptidyl-prolyl isomerases and photo-proline – containing client proteins. This chapter expands the photo-crosslinking non-canonical amino acids available for incorporation into proteins, and informs future efforts to apply photo-proline in studying protein-protein interactions.

4.3 Introduction

Protein-protein interactions (PPIs) are involved in a plethora of cellular processes and involve a wide range of protein interfaces and binding affinities.¹ One method among many² for studying protein-protein interaction partners involves photo-crosslinking non-canonical amino acids (ncAAs).³ These residues include a light-activatable moiety that, upon irradiation, generates a reactive intermediate capable of forming covalent bonds with nearby molecules. Photo-crosslinking ncAAs have been used, for instance, to identify protein-protein interactions,⁴ map interaction sites,⁵ and generate cross-linked protein biomaterials.⁶

Photo-crosslinking ncAAs are often incorporated into proteins in a site-specific manner. In these cases, the translational machinery is engineered such that a particular codon (traditionally the amber stop codon) instead encodes the ncAA of interest. *p*-Benzoylphenylalanine (Bpa) was first introduced as a crosslinking ncAA in a chemically synthesized peptide in 1986,⁷ then incorporated site-specifically into proteins using an engineered *Methanococcus jannaschii* tyrosyl-tRNA synthetase/tRNA pair in 2002.^{8,9} Other photo-crosslinking amino acids incorporated by nonsense suppression include *p*-azidophenylalanine (pAzF),¹⁰ 4'-[3-(trifluoromethyl)-3H-diazirin-3-yl]-1-phenylalanine

(TfmdPhe),¹¹ 3'-azibutyl-N-carbamoyl-lysine (AbK),¹² and 3-(3-methyl-3H-diazirine-3-yl)-propaminocarbonyl-N ϵ -L-lysine (DiZPK).⁴

Residue-specific replacement approaches have similarly led to the incorporation of photo-crosslinking ncAAs into proteins. These techniques rely upon promiscuity inherent to the endogenous translational machinery of the host organism: the amino acid analog replaces its canonical parent across the proteome. In particular, the diazirine-containing ncAAs photo-methionine, photo-leucine, and photo-lysine are accepted by the translational machinery of mammalian cells, and have been used to detect protein-protein interaction partners.^{13,14} Both photo-leucine and photo-methionine can also label proteins expressed in *E. coli*.^{15,16}

A similar photo-crosslinking amino acid derivative is photo-proline (photo-pro, Figure 4.1a), which includes a diazirine functionality at the C γ position of the pyrrolidine side chain. Photo-pro has been introduced into chemically-synthesized peptides¹⁷⁻²⁰ and small molecule inhibitors,²¹ and was first used to demonstrate the interaction between a peptidomimetic antibiotic and its target LptD.¹⁷ However, its incorporation into ribosomally-expressed proteins has not yet been reported.

Here, we demonstrate the incorporation of photo-pro into proteins expressed in *E. coli*. We show that photo-pro incorporation is enabled by high osmotic stress conditions or *proP* overexpression. We further describe our efforts to detect the substrates of peptidyl-prolyl isomerases (PPIases) in *E. coli* via photo-pro crosslinking; however, initial experiments

have not identified PPIase client proteins. Together, this chapter facilitates the production of recombinant proteins that include photo-pro residues, and may assist future efforts to apply this residue in studying protein-protein interactions *in vivo*.

4.4 Results and discussion

4.4.1 Incorporation of photo-proline

In Chapter II of this thesis, we described our efforts to identify an expanded set of proline analogs that could be incorporated into proteins expressed in *E. coli* (Figure 2.S1, Table 2.S1). We found that the non-canonical proline (ncPro) analog photo-pro could replace proline residues in recombinant proinsulin with an incorporation efficiency of 37% (Table 2.S1); *proS* overexpression and 0.3 M NaCl were used promote ncPro incorporation. In these initial incorporation studies, we also observed by mass spectrometry the presence of an ion corresponding to the loss of two protons from the proline-containing peptide (Figure 4.1b-c). A degradation product of photo-proline is 3,4-dehydroproline (dhp, Figure 4.1a, Figure 4.S1), so we concluded that this ion likely corresponds to the dhp-containing peptide. Because of its potential utility in photo-crosslinking studies, we sought to characterize and improve photo-pro incorporation.

The M157Q and C443G mutants of the *E. coli* prolyl-tRNA synthetase (ProRS) increase the incorporation of 4*S*-aminoproline and piperidine-2-carboxylic acid, respectively.^{22,23} However, overexpression of either mutant did not enhance photo-pro incorporation (Figure 4.1d, Table 4.1). We then screened photo-pro and NaCl concentrations for improved incorporation into proinsulin. Because proline is an osmoprotectant, high osmotic stress

conditions (such as high salt concentrations) increase expression of proline transporters²⁴ that enhance non-canonical proline (ncPro) uptake and incorporation.²³ Increasing the concentrations of both photo-pro and NaCl increased incorporation (Figure 4.1e-f, Table 4.1): 48% of peptides contained photo-pro when we increased the concentration of the ncPro to 1.0 mM (0.3 M NaCl); 0.5 M NaCl (0.3 mM photo-pro) increased incorporation further to 61%. No incorporation of photo-pro could be detected without added NaCl, highlighting the importance of amino acid uptake.

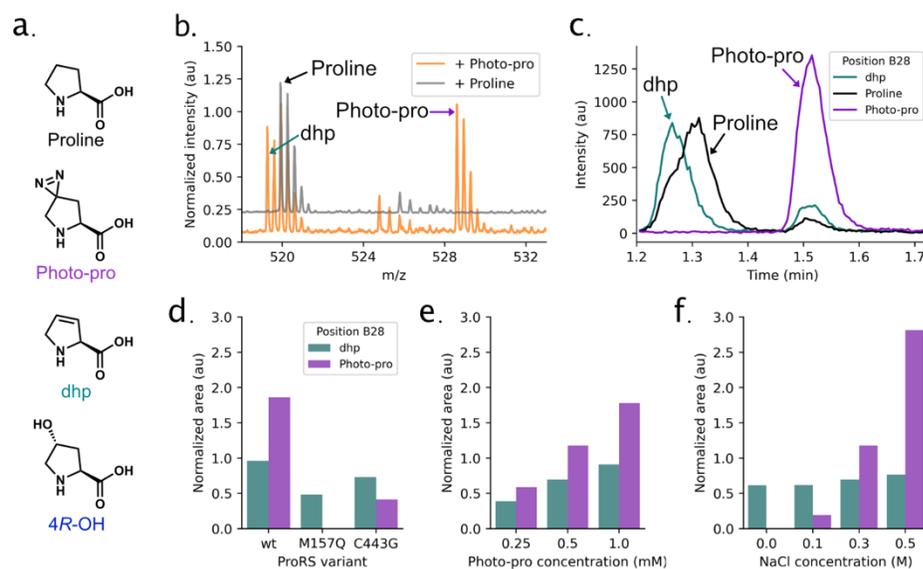


Figure 4.1. Improved incorporation of photo-pro. **a.** The chemical structures of proline and proline analogs described in this chapter. **b.** Mass spectra of the proinsulin peptide RGFFYTPKTRRE (obtained by Glu-C digestion; $[M+3H]^3$ ion) after proinsulin expression in the presence of proline or photo-pro. **c.** Extracted ion chromatograms (EIC; m/z = 520.3, proline; 519.3, dhp; 528.3, photo-pro; each corresponds to the $[M+3H]^3$ ion) of the +photo-pro sample shown in (b). **d-f.** Photo-pro and dhp incorporation vary across ProRS mutants (0.5 mM photo-pro, 0.3 M NaCl; d), photo-pro concentration (wild-type ProRS, 0.3 M NaCl; e), and NaCl concentration (wild-type ProRS, 0.5 mM photo-pro; f). Shown is the area under the curve corresponding to each ncPro peptide from the EIC, normalized to the proline-containing peptide in that sample.

Interestingly, these results also suggest that an appreciable level of dhp-containing peptide is the result of co-translational dhp incorporation, rather than photolysis after photo-pro incorporation. We observed the dhp peptide across all ProRS variants and NaCl concentrations tested, independent of photo-pro incorporation (Figure 4.1d,f). Its prevalence also increased with photo-pro concentration (Figure 4.1e), presumably due to the unavoidable correlation between concentrations of photo-pro and its thermal degradation products under these culturing conditions. We note that dhp is known to incorporate well into recombinant proteins in *E. coli*.²³ The low (~5%) levels of dhp produced due to thermal degradation over the course of a photo-pro labeling experiment (Figure 4.S1) appear to compete with photo-pro for incorporation at proline codons.

Table 4.1. Photo-pro incorporation efficiencies under osmotic stress conditions

Protein	ProRS	Photo-pro conc. (mM)	NaCl conc. (M)	Fraction proline	Fraction photo-pro	Fraction dhp
6xHis-PI	wt	0.5	0.3	0.26	0.49	0.25
6xHis-PI	M157Q	0.5	0.3	0.67	n.d.*	0.33
6xHis-PI	C443G	0.5	0.3	0.47	0.19	0.34
H27R-PI	wt	0.25	0.3	0.51	0.30	0.20
H27R-PI	wt	0.5	0.3	0.35	0.41	0.24
H27R-PI	wt	1	0.3	0.27	0.48	0.25
H27R-PI	wt	0.5	0	0.62	n.d.*	0.38
H27R-PI	wt	0.5	0.1	0.55	0.11	0.34
H27R-PI	wt	0.5	0.5	0.22	0.61	0.17

*n.d., not detected.

4.4.2 *ProP* overexpression leads to *ncPro* incorporation

High osmotic stress conditions necessary for photo-pro incorporation might interfere with *E. coli* cell physiology,²⁵ limiting the use of photo-pro to study protein-protein interactions

in vivo. We therefore pursued an alternative approach for photo-pro incorporation, namely the direct overexpression of the proline transporter *proP*²⁶ (Figure 4.2a-b). We found that expressing *proP* in proline medium, then inducing proinsulin expression after a medium shift into photo-pro medium, led to 73% photo-pro incorporation into proinsulin even in the absence of osmotic stress conditions. Using the *proP* overexpression approach, we obtained a 91% incorporation efficiency for the proline analog 4*R*-hydroxyproline (4*R*-OH, Figure 4.1a). These results are nearly identical to 4*R*-OH incorporation using osmotic stress conditions,²⁷ suggesting that *proP* overexpression is a general approach for ncPro uptake and incorporation (Figure 4.2c-d, Table 4.2).

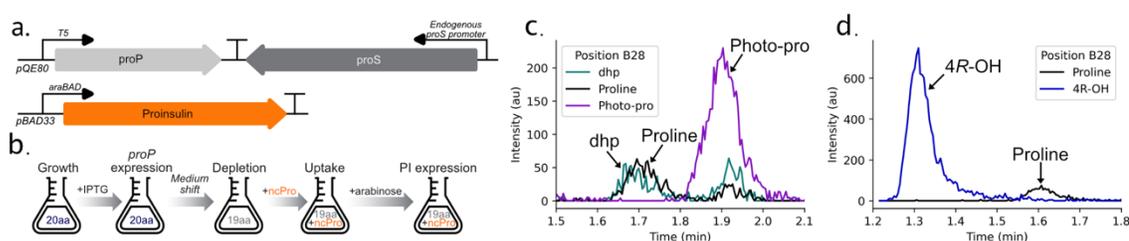


Figure 4.2. *proP* overexpression leads to proline analog incorporation. a-b. The constructs (a) and expression protocol (b) used for ncPro incorporation into recombinantly-expressed proinsulin by *proP* overexpression. c-d. Extracted ion chromatograms after photo-pro (c) and 4*R*-OH (d) incorporation. Proinsulins containing ncPro residues were digested with Glu-C and analyzed by LC-ESI-MS. The $[M+3H]^{+3}$ ion corresponding to the peptide RGFFYTPKTRRE was monitored ($m/z = 520.3$, proline; 519.3 , dhp; 528.3 , photo-pro; 525.3 , 4*R*-OH). Dhp and proline ions that co-eluted with the photo-pro ion (~ 1.9 min) were presumed to be degradation products of photo-proline formed during mass spectrometry.

Table 4.2. Proline analog incorporation efficiencies with *proP* overexpression

ncPro	ncPro conc. (mM)	Fraction proline	Fraction photo-pro	Fraction dhp	Fraction 4 <i>R</i> -OH
Photo-pro	1.0	0.135	0.743	0.122	–
4 <i>R</i> -OH	1.0	0.095	–	–	0.905

4.4.3 PPIase substrates were not detected with photo-proline

Peptidyl-prolyl isomerases (PPIases) are chaperones that promote amide *cis-trans* isomerization of proline residues. We wondered if the crosslinking capabilities of photo-pro could be leveraged to identify PPIase substrates in *E. coli*. Here, photo-pro was incorporated proteome-wide using the conditions that led to 74% incorporation into recombinant proinsulin (Figure 4.2), and samples irradiated to promote crosslinking at photo-pro positions. 3xFLAG-tagged PPIases, which was overexpressed (along with *proP*) in proline-containing medium prior to photo-pro labeling, was visualized by immunoblotting. We expected that client proteins could be enriched and identified by immunoprecipitation and mass spectrometry.

We were unable to detect the presence higher molecular weight species indicative of successful covalent crosslinking between a PPIase and its substrate by western blot (Figure 4.3a). Mass spectrometry was similarly unsuccessful: after PPIase immunoprecipitation, we did not observe an increase in the number of client proteins identified in the photo-pro treated sample, compared with the proline control (Figure 4.3b-c). Further, we could not confidently identify PPIase:client crosslinked peptides by mass spectrometry (data not shown).

Control photolysis experiments of solubilized photo-pro–proinsulin indicated 69-92% photoconversion by mass spectrometry after UV irradiation under similar conditions. We attempted to detect crosslinking to free, protected amino acids (N-boc and methyl ester

protected derivatives of cysteine, tyrosine, and glutamine); however, crosslinking at photo-pro positions was not identified (data not shown).

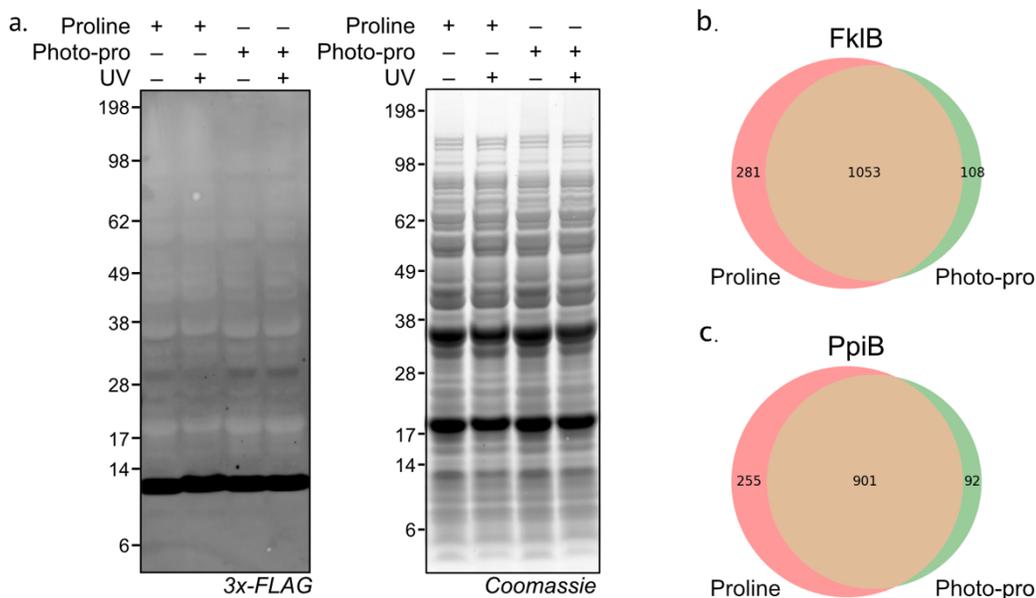


Figure 4.3. PPIase:client interactions were not detected using photo-proline. **a.** Western blot and SDS-PAGE analysis to detect PpiC crosslinked to its substrates. PpiC-3xFLAG and ProP were expressed in proline medium, then shifted to medium that instead contained photo-pro. Cells were irradiated (365 nm, 15 min), and proteins analyzed by SDS-PAGE. PpiC was detected using the 3xFLAG epitope; proline and -UV conditions were used as negative controls. No higher molecular weight species could be detected in +photo-pro/+UV conditions compared to the controls. Similar results were obtained for the eight other PPIases tested (not shown). **b, c.** FkIB-3xFLAG (b) and PpiB-3xFLAG (c) were expressed, possible substrates labeled with photo-pro, and cells irradiated as described in (a). Each PPIase, potentially cross-linked to its substrates by photo-pro, was enriched by immunoprecipitation. Samples were analyzed by bottom-up proteomics. We did not observe an increase in the number of hits for the samples labeled with photo-pro, compared to those labeled with proline.

Dissociation constants for PPIase:substrate interactions have been reported in the low micromolar to upper nanomolar range (for example, Ref. 28–30). This interaction is

perhaps too weak for sufficient crosslinking with photo-pro; we do note that successful photo-pro labeling was detected between two components with a binding constant in the tens of nanomolar range.²⁰ Perhaps carbene stabilization (for instance, by installing fluoro substituents about the pyrrolidine ring) would better promote cross-linking and substrate identification. The diazirine moiety, while small, might interrupt the interaction with the PPIase, or may interfere with client protein folding, solubility, and stability. Finally, optimized photolysis conditions, especially in the context of the cellular environment of *E. coli*, may improve crosslinking efficiency.

4.5 Conclusion

In conclusion, we have detailed the incorporation of the proline analog photo-pro into proteins expressed in *E. coli*. We find that photo-pro incorporation requires high osmotic stress conditions or *proP* overexpression. Despite obtaining good incorporation efficiencies in recombinantly-expressed proinsulin, we have thus far been unable to use photo-pro to identify PPIase substrates *in vivo*. Together, this work enables the production of recombinant proteins containing photo-pro at proline positions, and informs future efforts to use this residue in detecting protein-protein interactions *in vivo*.

4.6 Materials and methods

4.6.1 Chemicals

All chemicals were purchased from MilliporeSigma.

4.6.2 Enzymes

2x Q5 master mix (New England Biolabs) was used for all PCR applications. Gibson assembly was performed with the Repliq HiFi assembly mix from Quantabio.

4.6.3 Strains and plasmids

The proline-auxotrophic *E. coli* strain CAG18515 was obtained from the Coli Genetic Stock Center (CGSC) at Yale University. Strain DH10B was used for standard cloning operations; electrocompetent CAG18515 were transformed with purified plasmid products.

Plasmids pQE80PI-proS (which encodes 6xHis-PI and proS) and pQE80_H27R-PI_proS (which encodes H27R-PI and *proS*) were described in Chapter II of this thesis and references therein. Installation of the M157Q and C443G mutations into *proS* were also described in Chapter II.

4.6.4 Primers

AmpR-GA_fwd: GCAGTGTTATCACTCATGG

AmpR-GA_rev: GACAGTAAGAGAATTATGCAGTG

SLB7008_vec_rev: AGTTAATTTCTCCTCTTTAATGAATTC

SLB7008_vec_fwd: GGATCCGCATGCGAG

SLB7008_ins_fwd: GAATTCATTAAAGAGGAGAAATTA ACTATGCTGAAAAGGA
AAAAAGTAAAAC

SLB7008_ins_rev: CTCGCATGCGGATCCTTATTCATCAATTCGCGGATG

SLB7009_vec_fwd: GGTACCCGGGGATCCTCTAG

SLB7009_vec_rev: GAGCTCGAATTCGCTAGCCC

4.6.5 Nucleotide and amino acid sequences

The sequences of H27R-PI and *proS* are described in Chapter II of this thesis.

6xHis-PI:

ATGAGAGGATCGCATCACCATCACCATCACCGCTTTGTGAACCAGCACCTGT
GCGGTAGCCACCTGGTGAAGCTCTGTACCTGGTTTGCGGTGAGCGTGGTTTC
TTCTACACGCCaAAGACCCGCCGTGAAGCTGAAGATCTGCAGGTGGGCCAGG
TAGAACTGGGCGGTGGTCCGGGTGCCGGCTCTCTGCAACCGCTGGCACTGGA
AGGTTCCCTGCAAGCGCGTGGTATCGTAGAGCAGTGCTGTACTTCTATCTGCT
CCCTGTACCAGCTGGAGA ACTACTGTAATTAA

The sequence of the 6xHis leader peptide is underlined, proinsulin is in **bold**. The A-chain and B-chain in mature insulin are colored **red** and **blue**, respectively.

MRGSHHHHHHR**FVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQV**
GQVELGGGPGAGSLQPLALEGSLQARGIVEQCCTSICSLYQLENYCN

proP: The gene encoding *proP* is in UPPERCASE. The nucleotide sequence corresponding to the Gibson Assembly overlap regions used to install *proP* into plasmid pQE80_proP_proS are in lowercase.

gaattcattaaagaggagaaattaactATGCTGAAAAGGAAAAAAGTAAAACCGATTACCCTT
CGTGATGTCACCATTATTGATGACGGTAAACTGCGTAAAGCCATTACCGCAG
CATCACTGGGTAATGCAATGGAATGGTTCGATTTTGGTGTTTATGGTTTTGTT
GCTTACGCATTAGGTAAAGTTTTTTTCCCGGGGGCTGACCCAGCGTGCAGAT

GGTTGCTGCACTTGCCACTTTCTCCGTTCCCTTTCTGATTTCGACCGCTTGGCGG
ACTCTTCTTTGGTATGTTGGGCGATAAATATGGTCGCCAGAAGATCCTCGCTA
TCACTATTGTGATTATGTCGATCAGTACGTTCTGTATTGGCTTAATACCGTCCT
ACGACACGATTGGTATTTGGGCACCGATTCTGCTGTTGATCTGTAAGATGGCA
CAAGGTTTCTCGGTTCGGCGGTGAATATACCGGGGCGTCGATATTTGTTGCGG
AATACTCCCCTGACCGTAAACGTGGCTTTATGGGCAGCTGGCTGGACTTCGGT
TCTATTGCCGGGTTTGTGCTGGGTGCGGGCGTGGTGGTGTTAATTTTCGACCAT
TGTCGGCGAAGCGAACTTCCTCGATTGGGGCTGGCGTATTCCGTTCTTTATCG
CTCTGCCGTTAGGGATTATCGGGCTTTACCTGCGCCATGCGCTGGAAGAGACT
CCGGCGTTCCAGCAGCATGTCGATAAACTGGAACAGGGCGACCGTGAAGGTT
TGCAGGATGGCCCGAAAGTCTCGTTTAAAGAGATTGCCACTAAATACTGGCG
CAGCCTGTTGACATGTATTGGTCTGGTAATTGCCACCAACGTGACTTACTACA
TGTTGCTGACCTATATGCCGAGTTATTTGTGCGATAACCTGCATTACTCCGAA
GACCACGGGGTGCTGATTATTATCGCCATTATGATCGGTATGCTGTTTGTCCA
GCCGGTGATGGGCTTGCTGAGTGACCGTTTTGGCCGTCGTCCGTTTGTGCTAC
TTGGTAGTGTTGCCCTGTTTGTGTTGGCGATCCCGGGCGTTTATTCTGATTAACA
GTAACGTCATCGGCCTGATTTTTGCCGGGTTACTGATGCTGGCGGTGATCCTT
AACTGCTTTACGGGCGTTATGGCTTCTACCTTGCCAGCGATGTTCCCGACGCA
TATCCGTTACAGCGCGCTGGCGGCGGCATTTAATATTTCCGGTGCTGGTTGCCG
GTCTGACGCCAACGCTGGCGGCCTGGCTGGTCGAAAGCTCGCAGAATCTGAT
GATGCCTGCCTATTACCTGATGGTAGTGGCGGTGGTTGGTTTAATCACCGGCG
TAACCATGAAAGAGACGGCAAATCGTCCGTTGAAAGGTGCGACACCGGCGG
CGTCAGATATAACAGGAAGCGAAGGAAATTCTCGTCGAGCATTACGATAATAT
CGAGCAGAAAATCGATGATATTGACCACGAGATTGCCGATTTGCAGGCGAAA
CGTACCCGCCTGGTGCAGCAACATCCGCGAATTGATGAATAAggatccgcatgcgag
MLKRKKVKPITLRDVTIIDDGKLRKAITAASLGNAMEWFDGFGVYGFVAYALGKV
FFPGADPSVQMVAALATFSVPFLIRPLGGLFFGMLGDYGRQKILAITIVIMSISTF
CIGLIPSYDTIGIWAPILLICKMAQGFVSGGEYTGASIFVAEYSPDRKRGFMGSW
LDFGSIAGFVLGAGVVVLISTIVGEANFLDWGWRIPFFIALPLGIIGLYLRHALEET
PAFQQHVVDKLEQGDREGLQDGPVSKFKEIATKYWRSLLTCIGLVIATNVTYYYML

LTYMPSYLSHNLHYSEDHGVLIHAIMIGMLFVQPVMGLLSDRFGRRPFVLLGSVA
 LFVLAIPAFILINSNVIGLIFAGLLMLAVILNCFTGVMASLTPAMFPTHIRYSALAA
 AFNISVLVAGLTPTLAAWLVESSQNLMMPAYYLMVVAVVGLITGVTMKETANR
 PLKGATPAASDIQEAKEILVEHYDNIEQKIDDIDHEIADLQAKRTRLVQQHPRIDE

4.6.6 Cloning

pQE80_proP_proS: We installed the *proP* gene in place of proinsulin in plasmid pQE80_H27R-PI_proS using a three-part Gibson Assembly approach. The vector was amplified as two fragments which split the ampicillin resistance gene using the following primer sets: AmpR-GA_fwd / SLB7008_vec_rev & SLB7008_vec_fwd / AmpR-GA_rev. *proP* was amplified from the *E. coli* genome using colony PCR with primers SLB7008_ins_fwd & SLB7008_ins_rev.

pBAD33_PPIase-3xFLAG: Genes encoding nine *E. coli* proteins associated with PPIase activity (*fkIB*, *fkpA*, *fkpB*, *ppiA*, *ppiB*, *ppiC*, *slyD*, *surA*, and *tig*), with a C-terminal GS-linker and 3x-FLAG tag, were ordered as g-Block gene fragments from Integrated DNA technologies (IDT). They were installed into vector pBAD33 (amplified with primers SLB7009_vec_fwd & SLB7009_vec_rev) by Gibson Assembly. The amino acid sequences of each tagged PPIase are given in Table 4.S1.

4.6.7 Incorporation of photo-pro into proinsulin under osmotic stress conditions

A single colony of strain CAG18515 harboring plasmid pQE80_H27R-PI_proS or pQE80PI-proS was used to inoculate a culture of Luria Bertani (LB) medium supplemented with ampicillin. The culture was grown overnight at 37°C until stationary phase was

reached, then diluted 1:100 into 1x M9 medium, supplemented with all twenty amino acids. The medium composition of M9 is as follows: 8.5 mM NaCl, 18.7 mM NH₄Cl, 22 mM KH₂PO₄, 47.8 mM Na₂HPO₄, 0.1 mM CaCl₂, 1 mM MgSO₄, 3 mg L⁻¹ FeSO₄, 1 µg L⁻¹ trace metals [Cu²⁺, Mn²⁺, Zn²⁺, MoO₄²⁻], 35 mg L⁻¹ thiamine HCl, 10 mg L⁻¹ biotin, 20 mM D-glucose, 100 mg L⁻¹ ampicillin, 50 mg L⁻¹ of each L-amino acid.

The culture was grown at 37°C until it reached OD ~0.8, after which it was subjected to a medium shift: cells were pelleted via centrifugation (5 kg, 5 min, 4°C) and washed twice with ice-cold 0.9% NaCl. Washed cells were resuspended in 1.25x M9 –Pro, a 1.25x concentrated form of M9 that omits proline. The culture was split into 4 mL aliquots, and incubated for 30 min at 37°C to deplete residual proline. A 1 mL solution containing 1.25-5.0 mM ncPro and 0-2.5 M NaCl was added (0.25-1.0 mM ncPro and 0-0.5 M NaCl working concentrations). Cultures labeled with photo-pro were covered with aluminum foil to limit exposure to light. After 30 min of incubation at 37°C to allow for ncPro uptake, proinsulin expression was induced by the addition of 1 mM IPTG. Cultures were incubated for 2.5 h at 37°C, after which cells were harvested via centrifugation and stored at -80°C until further processing.

4.6.8 Incorporation of photo-pro into proinsulin with proP overexpression

A single colony of strain CAG18515 harboring plasmid pQE80_proP_proS and pBAD-PI was used to inoculate a culture of LB medium supplemented with ampicillin and chloramphenicol. The culture was grown overnight at 37°C until stationary phase was reached, then diluted 1:100 into 1x M9 medium, supplemented with all twenty amino acids.

The culture was grown at 37°C until it reached OD ~0.2, then 0.5 mM IPTG (to induce *proP* expression) was added. After 1.5 h incubation at 37°C, cells were pelleted via centrifugation (5 kg, 5 min, 4°C), washed twice with ice-cold 0.9% NaCl, then resuspended in 1x M9 –Pro (M9 medium as described above that lacks proline). The culture was incubated for 30 min at 37°C to deplete residual proline, after which 1 mM photo-pro, 4R-OH, or proline was added. Samples labeled with photo-pro were covered with aluminum foil to limit exposure to light. After 30 min incubation at 37°C to enable proline uptake, 0.1% arabinose was added to induce proinsulin expression. Cultures were incubated for 3 h at 37°C, then centrifuged and pellets stored at -80°C until further processing.

4.6.9 Incorporation of photo-pro with proP overexpression for PPIase crosslinking

A single colony of strain CAG18515 harboring plasmid pQE80_ proP_ proS and pBAD-PPIase was used to inoculate a culture of LB medium supplemented with ampicillin and chloramphenicol. Here, PPIase in pBAD-PPIase represents one of nine PPIases (Table 4.S1) with a C-terminal 3xFLAG tag. The culture was grown overnight at 37°C until stationary phase was reached, then diluted 1:100 into 1x M9 medium, supplemented with all twenty amino acids.

The culture was grown at 37°C until it reached OD ~0.2 (or ~0.4 for proteomics studies), then 0.5 mM IPTG (to induce *proP* expression) and 0.1% arabinose (to induce expression of the PPIase) were added. After 2 h incubation at 37°C, cells were pelleted via centrifugation (5 kg, 5 min, RT), washed twice with 1x M9 –Pro, then resuspended in 1x M9 –Pro. The culture was incubated for 30 min at 37°C to deplete residual proline, after

which 1 mM photo-pro or proline was added. Samples labeled with photo-pro were covered with aluminum foil to limit exposure to light. Cultures were incubated for 1.5 h at 37°C, then irradiated with UV light (see section 4.6.12 below).

4.6.10 Inclusion body isolation

Proinsulin is expressed in the inclusion body fraction in *E. coli*. Cell pellets were thawed and lysed with B-PER Complete (Thermo Fisher Scientific) for 1 h at room temperature with shaking, then centrifuged (14 kg, 10 min) and the supernatant discarded. The pellet (containing insoluble proinsulin) was washed once with Triton wash buffer (2 M urea, 20 mM Tris, 1% Triton X-100, pH 8.0), and twice with ddH₂O. The pellet was resuspended in solubilization buffer (8 M urea, 300 mM NaCl, 50 mM NaH₂PO₄, pH 8.0), and proinsulin was allowed to dissolve for 4.5 h at room temperature with shaking. Samples were centrifuged, and the supernatant removed for analysis (described in section 4.6.11 below).

4.6.11 LC-ESI-MS

To assess levels of photo-pro incorporation into proinsulins, samples were digested with Glu-C, which results in a peptide fragment containing ProB28 (RGFFYTPKTRRE). A 20 μ L aliquot of proinsulin was subjected to cysteine reduction (5 mM DTT, 55°C for 20 min) and alkylation (15 mM iodoacetamide, RT for 15 min in the dark), prior to 10-fold dilution into 100 mM NH₄HCO₃, pH 8.0 (100 μ L final volume). Digestion was started with addition of 0.6 μ L Glu-C (0.5 μ g μ L⁻¹ in ddH₂O) at 37°C for 2.5 h. The digestion reaction was

quenched by adding 10 μ L of 5% TFA. Peptides were analyzed by LC-ESI-MS at the Multi User Mass Spectrometry Laboratory at Caltech.

The $[M+3H]^{+3}$ ion corresponding to the peptide RGFFYTPKTRRE was monitored (m/z = 520.3, proline; 519.3, dhp; 528.3, photo-pro; 525.3, 4*R*-OH). Analog incorporation was calculated from the EIC by comparing the area under the curve (AUC) of the ncPro form of the peptide with the AUC of the canonical proline peptide.

Ions corresponding to dhp (m/z = 528.3) and proline (519.3) that co-eluted with the photo-pro ion were presumed to be degradation products formed during LC-MS and discarded from incorporation efficiency analysis. Further, the dhp peptide could not be resolved from the proline peptide. This led to an over-estimation of proline-peptide levels due to the presence of dhp-peptide isotopes at the same m/z as the proline-peptide. The incorporation efficiencies for photo-pro reported here thus represent under-estimated values.

4.6.12 UV irradiation

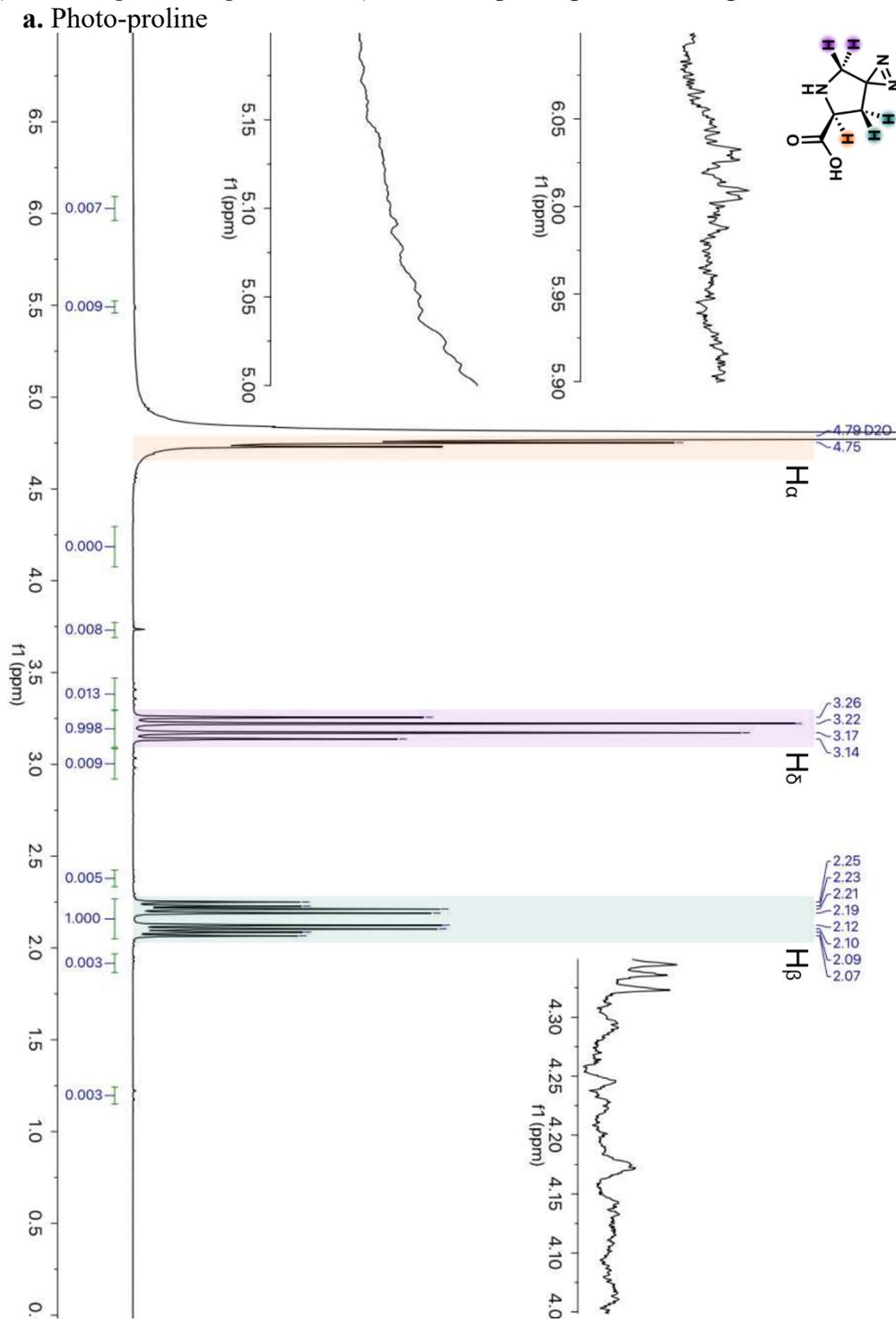
Cultures were aliquoted in 6-well tissue culture plates (5 mL culture per well) and irradiated for 15 min with 365 nm light (Black-Ray®XX-20BLB UV Bench Lamp, 20 Watt, P/N 95-0045-05, 10 cm distance) at 4 °C. Samples were then transferred to 15-50 mL tubes and centrifuged. Cell pellets were stored at -80°C until further processing and analysis.

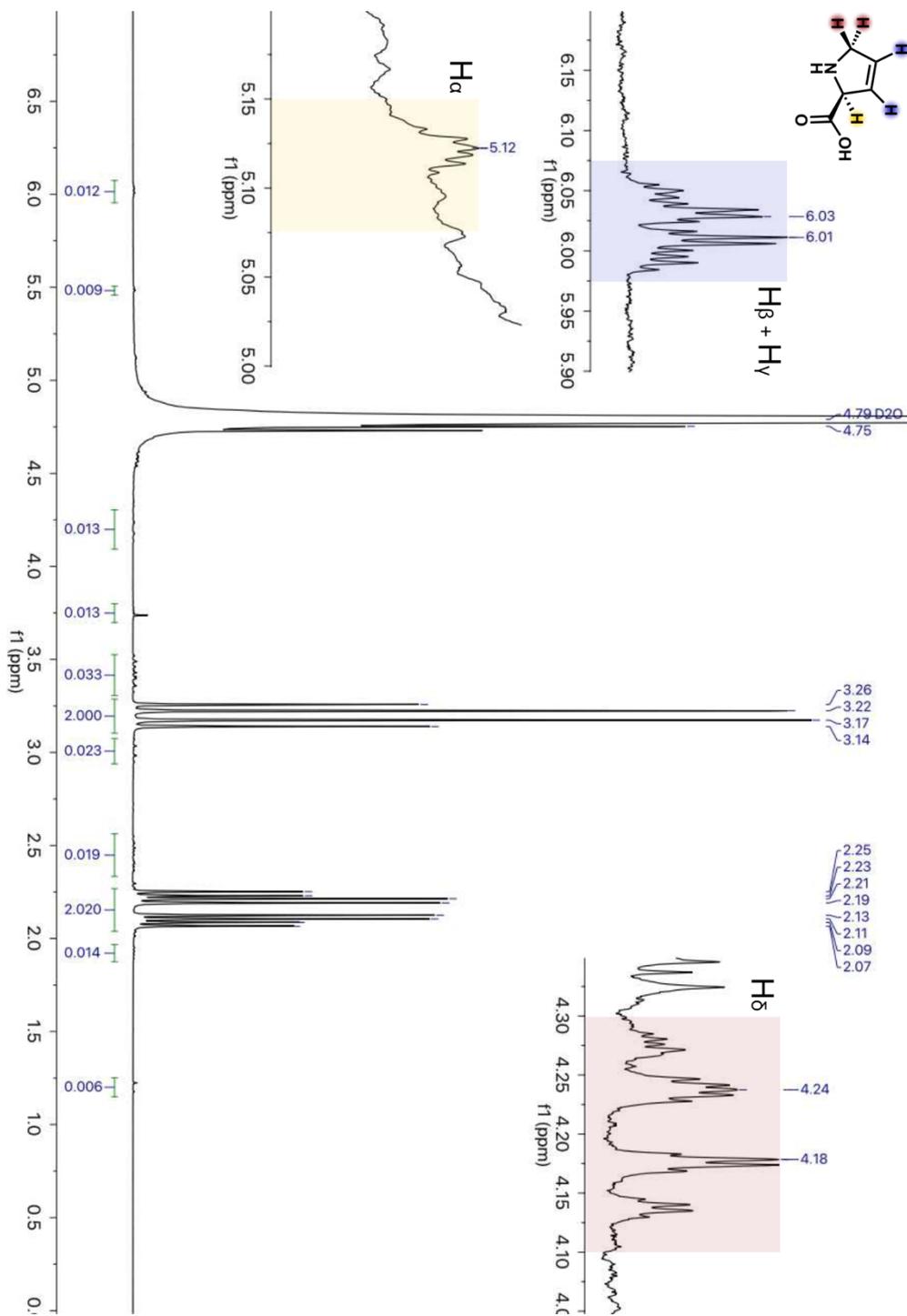
4.6.13 Western blot

Proteins separated by SDS-PAGE (NuPAGE 4-12% Bis-Tris, Thermo Fisher Scientific) were transferred onto a nitrocellulose membrane by an iBlot 2 gel transfer device (Thermo Fisher Scientific) using the manufacturer's protocols. Blots were blocked with 1 g dry milk in 20 mL PBST for 1 h at room temperature, then 4 μ L of a primary mouse M2 anti-FLAG antibody was added (1:5,000 dilution, 13 h, 4°C). The blot was washed four times with 20 mL PBST prior to the addition of the secondary antibody (goat anti-mouse-AF647 conjugate; 1:10,000 dilution, 2.5 h, RT). Blots were washed as before, then imaged on a Typhoon Trio (GE Healthcare) with a 633 nm laser and 670 nm bandpass filter.

4.7 Supplementary figures and tables

Figure 4.S1. Thermal degradation products of photo-pro include dhp. ^1H NMR spectra of photo-pro **a.** immediately after dissolution in D_2O ; **b.** after heating to 37°C for 3 h, dark (mimicking culturing conditions); **c.** heated photo-pro + 10% dhp.



b. Photo-proline; 37°C, 3 h

c. Photo-proline; 37°C, 3 h + 10% dhp

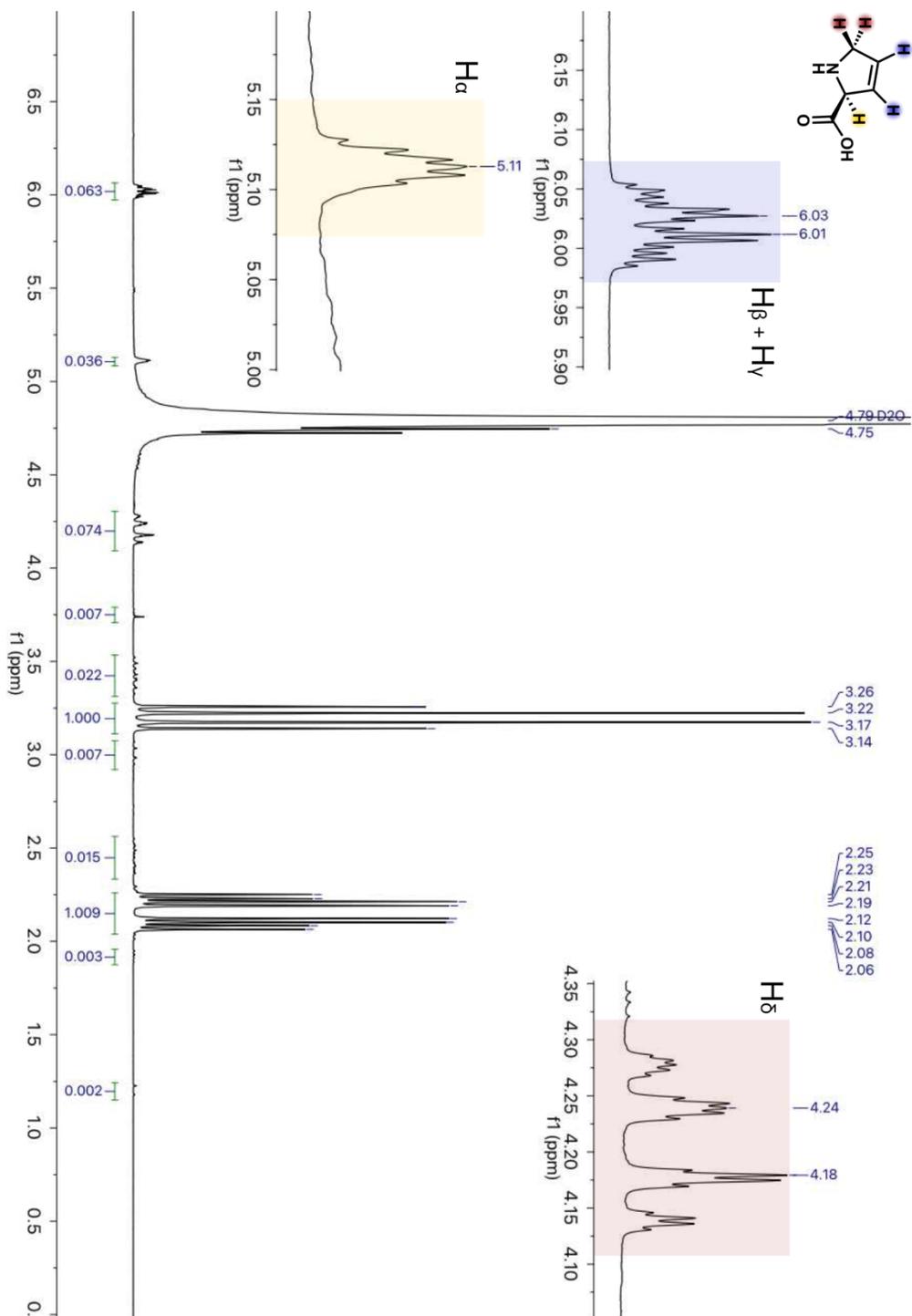


Table 4.S1. 3xFLAG-tagged PPIases

Name	Amino acid sequence*
fk1B	MTTPTFDTIEAQASYGIGLQVGGQQLSESGLEGLLPEALVAGIADALEG KHPAVPVDVVHRALREIHERADAVRRQRFQAMAAEGVKYLEENAK KEGVNSTESGLQFRVINQGEAIPARTDRVRVHYTGKLDGTVFDSSV ARGEPAEFPVNGVIPGWIEALTLMPVGSKWELTIPQELAYGERGAGAS IPPFSTLVFEVELLEILgsgsgsdykdhdgdykdhidykddddk
fkpA	MKSLFKVTLLATTMAVALHAPITFAAEAAKPATAADSKAAAFKNDDQ KSAYALGASLGRYMENSLKEQEKLGIKLDKDQLIAGVQDAFADKSKL SDQEIEQTLQAFEARVKSSAQAKMEKDAADNEAKGKEYREKFAKEK GVKTSSTGLVYQVVEAGKGEAPKDSDTVVVNYKGTLDGKEFDNSY TRGEPLSFRLDGVIPGWTEGLKNIKKGGKIKLVIPPELAYGKAGVPGIP PNSTLVFDVELLDVKPAPKADAKPEADAKAADSACKgsgsgsdykdhdgd ykdhidykddddk
fkpB	MSESVQSNSAVLVHFTLKLDDGTTAESTRNNGKPALFRLGDASLSEG LEQHLLGLKVGDKTTFSLEPDAAFGVSPDLIQYFSRREFMDAGEPEI GAIMLFTAMDGSEMPGVIREINGDSITVDFNHPLAGQTVHFVDFIEVLEID PALEAgsgsgsdykdhdgdykdhidykddddk
ppiA	MFKSTLAAMAAVFALSALSPAAMAAKGDPHVLLTTSAGNIELELDK QKAPVSVQNFVDYVNSGFYNNTTFHRVIPGFMIQGGGFTEQMQQKKP NPIKNEADNGLRNTRGTIAMARTADKDSATSQFFINVADNAFLDHG QRDFGYAVFGKVVKGMDVADKISQVPTHVGPYQNVPSKPVVILSA KVLPgsgsgsdykdhdgdykdhidykddddk
ppiB	MVTFHTNHGDIVIKTFDDKAPETVKNFLDYCREGFYNNTIFHRVINGF MIQGGGFEPGMKQKATKEPIKNEANGLKNTRGTLAMARTQAPHS TAQFFINVVDNDFLNFSGESLQGWGYCVFAEVVDGMDVVDKIKGVA TGRSGMHQDVPKEDVIIESVTVSEgsgsgsdykdhdgdykdhidykddddk
ppiC	MAKTAAALHILVKEEKLALDLLEQIKNGADFGKLAKKHSICPSGKRG GDLGEFRQGMVPAFDKVVVFSCPVLEPTGPLHTQFGYHIIKVLRYNgg sggsdykdhdgdykdhidykddddk
slyD	MKVAKDLVVSLAYQVRTEDGVLVDESPVSAPLDYLHGHGSLISGLET ALEGHEVGDKFDVAVGANDAYGQYDENLVQRPKDVFMGVDELQV GMRFLAETDQGPVPEITAVEDDHVVVDGNHMLAGQNLKFNVEVV AIREATEEELAHGHVHGAHDHHDHHDGCGGHHGHHDHGHEHGGE GCCGGKNGGCGCHgsgsgsdykdhdgdykdhidykddddk

surA MKNWKTL L L G I A M I A N T S F A A P Q V V D K V A A V V N N G V V L E S D V D G L
M Q S V K L N A A Q A R Q Q L P D D A T L R H Q I M E R L I M D Q I I L Q M G Q K M G V K I
S D E Q L D Q A I A N I A K Q N N M T L D Q M R S R L A Y D G L N Y N T Y R N Q I R K E M I
I S E V R N N E V R R R I T I L P Q E V E S L A Q Q V G N Q N D A S T E L N L S H I L I P L P E N
P T S D Q V N E A E S Q A R A I V D Q A R N G A D F G K L A I A H S A D Q Q A L N G G Q M
G W G R I Q E L P G I F A Q A L S T A K K G D I V G P I R S G V G F H I L K V N D L R G E S K N
I S V T E V H A R H I L L K P S P I M T D E Q A R V K L E Q I A A D I K S G K T T F A A A A K E
F S Q D P G S A N Q G G D L G W A T P D I F D P A F R D A L T R L N K G Q M S A P V H S S F
G W H L I E L L D T R N V D K T D A A Q K D R A Y R M L M N R K F S E E A A S W M Q E Q
R A S A Y V K I L S N g g s g g s d y k d h d g d y k d h d i d y k d d d d k

tig M Q V S V E T T Q G L G R R V T I T I A A D S I E T A V K S E L V N V A K K V R I D G F R K G
K V P M N I V A Q R Y G A S V R Q D V L G D L M S R N F I D A I I K E K I N P A G A P T Y V P
G E Y K L G E D F T Y S V E F E V Y P E V E L Q G L E A I E V E K P I V E V T D A D V D G M L
D T L R K Q Q A T W K E K D G A V E A E D R V T I D F T G S V D G E E F E G G K A S D F V L
A M G Q G R M I P G F E D G I K G H K A G E E F T I D V T F P E E Y H A E N L K G K A A K F
A I N L K K V E E R E L P E L T A E F I K R F G V E D G S V E G L R A E V R K N M E R E L K S
A I R N R V K S Q A I E G L V K A N D I D V P A A L I D S E I D V L R R Q A A Q R F G G N E K
Q A L E L P R E L F E E Q A K R R V V V G L L L G E V I R T N E L K A D E E R V K G L I E E M
A S A Y E D P K E V I E F Y S K N K E L M D N M R N V A L E E Q A V E A V L A K A K V T E
K E T T F N E L M N Q Q A g g s g g s d y k d h d g d y k d h d i d y k d d d d k

* Amino acid sequences in UPPERCASE correspond to the indicated PPIase; those in lowercase represent the C-terminal linker and 3xFLAG tag.

4.8 References

- (1) Veselovsky, A. V.; Ivanov, Y. D.; Ivanov, A. S.; Archakov, A. I.; Lewi, P.; Janssen, P. Protein–Protein Interactions: Mechanisms and Modification by Drugs. *J Mol Recognit* **2002**, *15* (6), 405–422. <https://doi.org/10.1002/JMR.597>.
- (2) Zhou, M.; Li, Q.; Wang, R. Current Experimental Methods for Characterizing Protein-Protein Interactions. *ChemMedChem* **2016**, *11* (8), 738–756. <https://doi.org/10.1002/cmdc.201500495>.
- (3) Yang, Y.; Song, H.; Chen, P. R. Genetically Encoded Photocrosslinkers for Identifying and Mapping Protein-Protein Interactions in Living Cells. *IUBMB Life* **2016**, *68* (11), 879–886. <https://doi.org/10.1002/iub.1560>.
- (4) Zhang, M.; Lin, S.; Song, X.; Liu, J.; Fu, Y.; Ge, X.; Fu, X.; Chang, Z.; Chen, P. R. A Genetically Incorporated Crosslinker Reveals Chaperone Cooperation in Acid Resistance. *Nat Chem Biol* **2011**, *7* (10), 671–677. <https://doi.org/10.1038/nchembio.644>.
- (5) Mori, H.; Ito, K. Different Modes of SecY-SecA Interactions Revealed by Site-Directed in Vivo Photo-Cross-Linking. *Proc Natl Acad Sci* **2006**, *103* (44), 16159–16164. <https://doi.org/10.1073/pnas.0606390103>.
- (6) Carrico, I. S.; Maskarinec, S. A.; Heilshorn, S. C.; Mock, M. L.; Liu, J. C.; Nowatzki, P. J.; Franck, C.; Ravichandran, G.; Tirrell, D. A. Lithographic Patterning of Photoreactive Cell-Adhesive Proteins. *J Am Chem Soc* **2007**, *129*, 4874–4875. <https://doi.org/10.1021/ja070200b>.
- (7) Kauers, J. C.; Erickson-Viitanen, S.; Wolfe, H. R.; Degrados, W. F. P-Benzoyl-L-Phenylalanine, A New Photoreactive Amino Acid. *J Biol Chem* **1986**, *261* (23), 10695–10700. [https://doi.org/10.1016/S0021-9258\(18\)67441-1](https://doi.org/10.1016/S0021-9258(18)67441-1).
- (8) Chin, J. W.; Martin, A. B.; King, D. S.; Wang, L.; Schultz, P. G. Addition of a Photocrosslinking Amino Acid to the Genetic Code of *Escherichia coli*. *Proc Natl Acad Sci* **2002**, *99* (17), 11020–11024. <https://doi.org/10.1073/pnas.172226299>.
- (9) Chin, J. W.; Schultz, P. G. In Vivo Photocrosslinking with Unnatural Amino Acid Mutagenesis. *ChemBioChem* **2002**, *3* (11), 1135–1137. <https://doi.org/10.1002/1439-7633>.
- (10) Chin, J. W.; Santoro, S. W.; Martin, A. B.; King, D. S.; Wang, L.; Schultz, P. G. Addition of P-Azido-L-Phenylalanine to the Genetic Code of *Escherichia coli*. *J Am Chem Soc* **2002**, *124* (31), 9026–9027. <https://doi.org/10.1021/ja027007w>.
- (11) Tippmann, E. M.; Liu, W.; Summerer, D.; Mack, A. V.; Schultz, P. G. A Genetically Encoded Diazirine Photocrosslinker in *Escherichia coli*. *ChemBioChem* **2007**, *8* (18), 2210–2214. <https://doi.org/10.1002/cbic.200700460>.
- (12) Ai, H.; Shen, W.; Sagi, A.; Chen, P. R.; Schultz, P. G. Probing Protein-Protein Interactions with a Genetically Encoded Photo-Crosslinking Amino Acid. *ChemBioChem* **2011**, *12* (12), 1854–1857. <https://doi.org/10.1002/cbic.201100194>.

- (13) Suchanek, M.; Radzikowska, A.; Thiele, C. Photo-Leucine and Photo-Methionine Allow Identification of Protein-Protein Interactions in Living Cells. *Nat Methods* **2005**, *2* (4), 261–267. <https://doi.org/10.1038/nmeth752>.
- (14) Yang, T.; Li, X. M.; Bao, X.; Fung, Y. M. E.; Li, X. D. Photo-Lysine Captures Proteins That Bind Lysine Post-Translational Modifications. *Nat Chem Biol* **2016**, *12* (2), 70–72. <https://doi.org/10.1038/nchembio.1990>.
- (15) Black, D.; Tran, Q.-K.; Keightley, A.; Chinawalkar, A.; McMullin, C.; Persechini, A. Evaluating Calmodulin–Protein Interactions by Rapid Photoactivated Cross-Linking in Live Cells Metabolically Labeled with Photo-Methionine. *J Proteome Res* **2019**, *18* (10), 3780–3791. <https://doi.org/10.1021/acs.jproteome.9B00510>.
- (16) MacKinnon, A. L.; Garrison, J. L.; Hegde, R. S.; Taunton, J. Photo-Leucine Incorporation Reveals the Target of a Cyclodepsipeptide Inhibitor of Cotranslational Translocation. *J Am Chem Soc* **2007**, *129* (47), 14560–14561. <https://doi.org/10.1021/ja076250y>.
- (17) Srinivas, N.; Jetter, P.; Ueberbacher, B. J.; Werneburg, M.; Zerbe, K.; Steinmann, J.; Van Der Meijden, B.; Bernardini, F.; Lederer, A.; Dias, R. L. A.; Misson, P. E.; Henze, H.; Zumbunn, J.; Gombert, F. O.; Obrecht, D.; Hunziker, P.; Schauer, S.; Ziegler, U.; Käch, A.; Eberl, L.; Riedel, K.; Demarco, S. J.; Robinson, J. A. Peptidomimetic Antibiotics Target Outer-Membrane Biogenesis in *Pseudomonas Aeruginosa*. *Science (1979)* **2010**, *327* (5968), 1010–1013. <https://doi.org/10.1126/science.1182749>.
- (18) Urfer, M.; Bogdanovic, J.; Monte, F. Lo; Moehle, K.; Zerbe, K.; Omasits, U.; Ahrens, C. H.; Pessi, G.; Eberl, L.; Robinson, J. A. A Peptidomimetic Antibiotic Targets Outer Membrane Proteins and Disrupts Selectively the Outer Membrane in *Escherichia coli*. *J Biol Chem* **2016**, *291* (4), 1921–1932. <https://doi.org/10.1074/jbc.M115.691725>.
- (19) Vetterli, S. U.; Zerbe, K.; Müller, M.; Urfer, M.; Mondal, M.; Wang, S. Y.; Moehle, K.; Zerbe, O.; Vitale, A.; Pessi, G.; Eberl, L.; Wollscheid, B.; Robinson, J. A. Thanatin Targets the Intermembrane Protein Complex Required for Lipopolysaccharide Transport in *Escherichia coli*. *Sci Adv* **2018**, *4* (11), eaau2634. <https://doi.org/10.1126/sciadv.aau2634>.
- (20) Andolina, G.; Bencze, L. C.; Zerbe, K.; Müller, M.; Steinmann, J.; Kocherla, H.; Mondal, M.; Sobek, J.; Moehle, K.; Malojčić, G.; Wollscheid, B.; Robinson, J. A. A Peptidomimetic Antibiotic Interacts with the Periplasmic Domain of LptD from *Pseudomonas Aeruginosa*. *ACS Chem Biol* **2018**, *13* (3), 666–675. <https://doi.org/10.1021/acschembio.7B00822>.
- (21) Zhou, Z.; Hu, T.; Zhou, X.; Wildum, S.; Garcia-Alcalde, F.; Xu, Z.; Wu, D.; Mao, Y.; Tian, X.; Zhou, Y.; Shen, F.; Zhang, Z.; Tang, G.; Najera, I.; Yang, G.; Shen, H. C.; Young, J. A. T.; Qin, N. Heteroaryldihydropyrimidine (HAP) and Sulfamoylbenzamide (SBA) Inhibit Hepatitis B Virus Replication by Different Molecular Mechanisms. *Sci Rep* **2017**, *7* (1), 1–12. <https://doi.org/10.1038/srep42374>.
- (22) Fang, K. Y. Modulating Biophysical Properties of Insulin with Non-Canonical Mutagenesis at Position B28, California Institute of Technology, 2017. <https://doi.org/10.7907/Z9V40S6J>.

- (23) Kim, W.; George, A.; Evans, M.; Conticello, V. P. Cotranslational Incorporation of a Structurally Diverse Series of Proline Analogues in an *Escherichia coli* Expression System. *ChemBioChem* **2004**, *5* (7), 928–936. <https://doi.org/10.1002/cbic.200400052>.
- (24) Grothe, S.; Krogsrud, R. L.; McClellan, D. J.; Milner, J. L.; Wood, J. M. Proline Transport and Osmotic Stress Response in *Escherichia coli* K-12. *J Bacteriol* **1986**, *166* (1), 253–259. <https://doi.org/10.1128/jb.166.1.253-259.1986>.
- (25) Weber, A.; Kögl, S. A.; Jung, K. Time-Dependent Proteome Alterations under Osmotic Stress during Aerobic and Anaerobic Growth in *Escherichia coli*. *J Bacteriol* **2006**, *188* (20), 7165–7175. <https://doi.org/10.1128/jb.00508-06>.
- (26) Sasaki, H.; Sato, D.; Oshima, A. Importance of the High-Expression of Proline Transporter PutP to the Adaptation of *Escherichia coli* to High Salinity. *Biocontrol Sci* **2017**, *22* (2), 121124. <https://doi.org/10.4265/bio.22.121>.
- (27) Lieblisch, S. A.; Fang, K. Y.; Cahn, J. K. B.; Rawson, J.; LeBon, J.; Teresa Ku, H.; Tirrell, D. A. 4S-Hydroxylation of Insulin at ProB28 Accelerates Hexamer Dissociation and Delays Fibrillation. *J Am Chem Soc* **2017**, *139*, 8384–8387. <https://doi.org/10.1021/jacs.7b00794>.
- (28) Kovermann, M.; Zierold, R.; Haupt, C.; Löw, C.; Balbach, J. NMR Relaxation Unravels Interdomain Crosstalk of the Two Domain Prolyl Isomerase and Chaperone SlyD. *Biochim Biophys Acta Proteins Proteom* **2011**, *1814* (7), 873–881. <https://doi.org/10.1016/j.bbapap.2011.03.016>.
- (29) Saio, T.; Guan, X.; Rossi, P.; Economou, A.; Kalodimos, C. G. Structural Basis for Protein Antiaggregation Activity of the Trigger Factor Chaperone. *Science (1979)* **2014**, *344* (6184). <https://doi.org/10.1126/science.1250494>.
- (30) Quistgaard, E. M.; Weininger, U.; Ural-Blimke, Y.; Modig, K.; Nordlund, P.; Akke, M.; Löw, C. Molecular Insights into Substrate Recognition and Catalytic Mechanism of the Chaperone and FKBP Peptidyl-Prolyl Isomerase SlyD. *BMC Biol* **2016**, *14* (1), 1–25. <https://doi.org/10.1186/s12915-016-0300-3>.

TOWARD EXPANDING NON-CANONICAL PROLINE INCORPORATION BY PROLYL-TRNA SYNTHETASE ENGINEERING

5.1 Contributions

Bradley R. Silverman assisted with flow cytometry, FACS, and microscopy; and provided helpful discussions. Katharine Y. Fang initiated the split-GFP approach and performed the cloning of those initial constructs and the first error-prone *proS* library. Alejandro Lopez contributed to cloning M157Q and C443G *proS* mutants. Stephanie L. Breunig prepared all other constructs, performed all experiments, and analyzed all data.

5.2 Abstract

The promiscuity of the translational machinery of *E. coli* has enabled the incorporation of chemically diverse proline analogs into recombinant proteins, and point mutants of the prolyl-tRNA synthetase (ProRS) can increase the residue-specific incorporation for several non-canonical proline residues. However, robust high-throughput methods are needed to identify ProRS variants capable of accepting more chemically diverse proline analogs. Here, we discuss several attempts to develop a screening platform for the directed evolution of the *E. coli* ProRS. First, we describe our efforts involving full-length and split fluorescent protein reporters. Next, we touch on alternative approaches that seek to detect inclusion bodies, or proline analog toxicity. Finally, we detail an approach that relies upon specific labeling of a short tetracysteine motif by the small molecule fluorophore FLAsH.

These efforts highlight the difficulty in engineering the *E. coli* ProRS for improved proline analog incorporation, and lay the groundwork for future ProRS engineering endeavors.

5.3 Introduction

The promiscuity within the translational machinery of *E. coli* has enabled the incorporation of chemically diverse non-canonical amino acids (ncAAs) into proteins in living cells; protein engineering efforts have further expanded ncAA incorporation technologies. In most cases, the gatekeepers of translational fidelity are the aminoacyl-tRNA synthetases (aaRSs),¹ so increased ncAA incorporation is often achieved by aaRS engineering. Amber suppression and related site-specific incorporation approaches require the development of an orthogonal aaRS/tRNA pair.² While not required in all cases, residue-specific incorporation has similarly been assisted by aaRS engineering. For instance, separately engineered MetRS variants have enabled the incorporation of the methionine analogs trifluoronorleucine (Tfn)³ and azidonorleucine (Anl).⁴

To date, a residue-specific incorporation approach is the only method to incorporate proline analogs into recombinant proteins. Many non-canonical proline (ncPro) residues are readily accepted by the endogenous *E. coli* translational machinery,⁵ or simply require overexpression of the wild-type prolyl-tRNA synthetase (ProRS), and high NaCl concentrations to promote proline uptake.⁶ Point mutations in the ProRS active site have also been reported to improve ncPro incorporation.⁷ However, the chemical space about proline is vast, and many ncPro residues of interest cannot be incorporated efficiently into recombinant proteins (for example, see Figure 2.S1 and Table 2.S1).

Directed evolution is a powerful technology to engineer enzymes for applications beyond their endogenous activities;⁸ we note numerous examples of aaRS engineering for ncAA incorporation.² However, these successes have not involved all the aminoacyl-tRNA synthetases used by biology. Commonly evolved aaRS/tRNA pairs include the PylRS/tRNA_{CUA} pairs from *Methanosarcina barkeri* and *Methanosarcina mazei*, and the TyrRS/tRNA_{CUA} pairs from *Escherichia coli* and *Methanocaldococcus janaschii*.^{2,9} The sole report in the literature that describes ProRS engineering resulted in orthogonal ProRS/tRNA^{Pro} pairs,¹⁰ but no site-specific incorporation of proline analogs was reported. Later efforts to expand upon this work¹¹ point to the difficulties in site-specific ncPro incorporation.

We sought to engineer the *E. coli* ProRS to accept a diverse range of proline analogs, with the goal of incorporating these residues into recombinant proteins in a residue-specific¹² (rather than site-specific) manner. In this chapter, we describe several of these efforts. Most designs are fluorescence-based platforms that might be used to sort large ProRS libraries by fluorescence-activated cell sorting (FACS). We explored a split-GFP system and a FIAsh-based labeling protocol to measure ncPro incorporation. Major challenges encountered with the split-GFP approach include reporter protein solubility and limited dynamic range. FIAsh labeling provided good analytical measurements for ncPro incorporation, but we faced significant hurdles to effectively sort ProRS libraries by FACS. We also describe a few alternative approaches that were similarly unsuccessful. Although we have not yet identified ProRS mutants capable of improved ncPro incorporation by high throughput screening to date, this work informs future ProRS engineering efforts.

5.4 Results and discussion

5.4.1 Full-length fluorescent proteins are poor reporters of proline analog incorporation

In FACS-based screening methods, GFP (or another full-length fluorescent protein) is often used as a reporter for ncAA incorporation. The codon(s) encoding the ncAA of interest are placed at a permissive location in the fluorescent protein, and codon readthrough (presumably by successful translation with the ncAA) leads to complete protein synthesis and successful chromophore formation.

The unique properties of proline, and diversity in the chemical and conformational properties of proline analogs, complicate a fluorescent protein approach for ProRS engineering. Replacing proline with ncPro residues in a full-length protein often alters protein folding and stability, as discussed in Chapter I of this thesis. In fact, global ncPro replacement in fluorescent proteins can lead to unpredictable results: global replacement of proline with 4*S*-fluoroproline (4*S*-F) leads to a fluorescent GFP variant with faster folding kinetics than its proline-containing parent. However, introduction of its diastereomer 4*R*-fluoroproline (4*R*-F) results in an insoluble GFP.¹³ Opposite stereospecific effects were obtained with mRFP1; in that case, 4*S*-F incorporation led to an insoluble protein, while 4*R*-F accelerated protein folding.¹⁴

A more recent report replaced the proline residues in three fluorescent proteins (EGFP, NowGFP, and KillerOrange) with 4*R*-F, 4*S*-F, 4,4-difluoroproline (44-diF) and 3,4-dehydroproline (dhp).¹⁵ In all cases, replacement with 4*S*-F or dhp was tolerated, while 4*R*-F and 44-diF did not lead to fluorescence. For examples in which ncPro and fluorescent protein pairs led to a mature fluorescent protein, the authors measured refolding kinetics

after chemical denaturation. In the context of EGFP, incorporation of 4S-F sped refolding, while dhp slowed it. Interestingly, all variants of NowGFP and KillerOrange did not refold once denatured.¹⁵

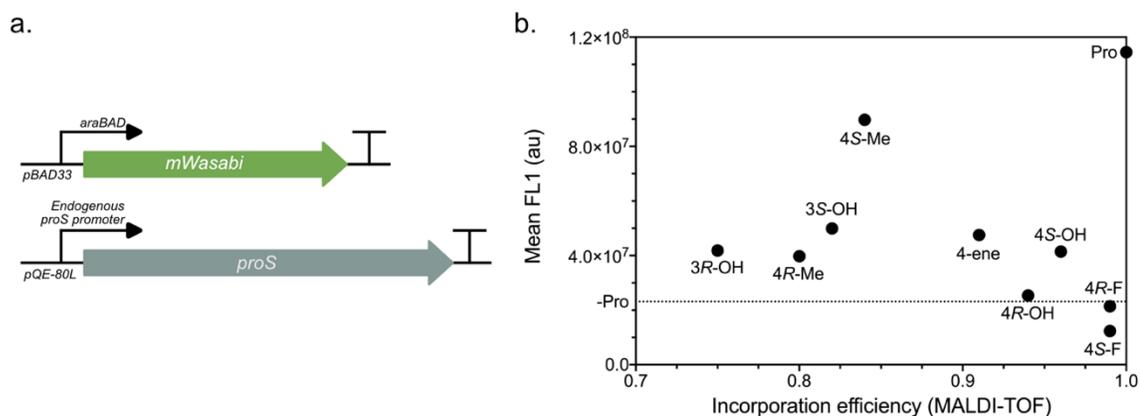


Figure 5.1. mWasabi fluorescence does not correspond to proline analog incorporation efficiencies. **a.** mWasabi expression was placed under control of the arabinose-inducible *araBAD* promoter, and *proS* was overexpressed from *pQE-80L* backbone. **b.** mWasabi fluorescence (FL1 by flow cytometry) as a function of incorporation efficiency (as determined by MALDI-TOF). mWasabi was expressed for 2.5 h at 37 °C under ncPro-incorporation conditions, and the resulting fluorescence was assessed by flow cytometry. Incorporation efficiency was determined by MALDI-TOF mass spectrometry after expression of ncPro-containing proinsulins.

Similarly unpredictable behavior was observed in our hands. We expressed mWasabi in *E. coli* under ncPro-incorporation conditions (Figure 5.1a), and assessed fluorescence by flow cytometry. We found that fluorescence is not a good predictor of ncPro incorporation efficiency, as determined by MALDI-TOF (Figure 5.1b). In fact, in some cases (such as 4S-F), the measured fluorescence is *less than that of a sample in which no proline has been added* (dotted line), despite the fact that these residues are known to have high incorporation efficiencies in *E. coli*.⁵ We suggest that 4S-F out-competes residual proline for incorporation into mWasabi, but leads to an improperly folded protein. The result is a

sample that is less fluorescent than the “–Proline” negative control. We also observe that known conformational preferences of the proline analogs tested¹⁶ do not correlate with fluorescence; for instance, 4*S*-Me and 4*R*-OH both prefer the *exo* ring pucker and *trans* amide isomer compared to other proline analogs assessed,^{16,17} yet lead to divergent fluorescent behaviors when incorporated into mWasabi (Figure 5.1b).

This unpredictable fluorescence prohibits the use of full-length fluorescent proteins as a general screening output for ncPro incorporation. In the directed evolution of the MetRS for Tfn incorporation, methionine residues were first removed from GFP, then reintroduced at permissive locations.³ We anticipated difficulties in applying a similar approach to obtaining a proline-free fluorescent protein due to the imino acid’s unique properties. Further, Pro89 (which exists as the *cis* isomer) and Pro196 are among the most widely conserved residues in GFPs and GFP-like proteins,¹⁸ so changes at those positions will not likely be tolerated. To the best of our knowledge, no reported fluorescent protein contains fewer than two proline residues, and proteins with few proline residues bind exogenous co-factors. We concluded that full-length fluorescent proteins are unable to act as a generalizable, robust approach to determining ncPro incorporation levels, and therefore pursued alternative designs.

5.4.2 *Split-GFP as a reporter for proline analog incorporation*

The presence of ncPro residues within any reporter protein will likely interfere with the function of that reporter. A split-GFP approach, in which one GFP fragment does not contain proline residues, might circumvent these issues.¹⁹ In this case, a larger, proline-containing unit (here, GFP1-10) would be expressed in proline-containing medium. Proline codon(s) would be placed N-terminal to the proline-free fragment (GFP11), which would

be expressed in ncPro medium after a medium shift. As a result, expression of the second, proline-free fragment would depend upon readthrough of the proline codons, yet the fluorescent output should not be affected by the conformational properties of a particular ncPro (Figure 5.2a).

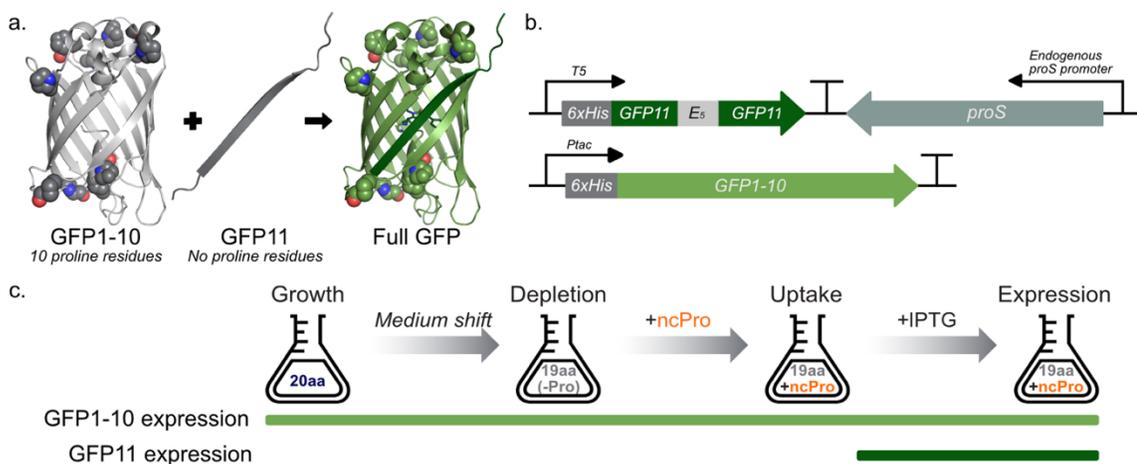


Figure 5.2. Design of split-GFP to measure proline analog incorporation. **a.** GFP1-10 (which contains ten proline residues, rendered as spheres) is expressed in proline-containing medium. Proline-dependent expression of the GFP11 fragment is induced after a medium shift into ncPro-containing medium. Complementation of the two fragments enables chromophore formation, providing a fluorescent output for ncPro incorporation (PDB: 2B3P). **b.** The initial construct design, as previously described.¹⁹ **c.** Experimental expression scheme for measuring ncPro incorporation by GFP1-10/11 complementation. GFP1-10 is under the control of a strong constitutive promoter, so is expressed throughout the duration of the experiment; expression of GFP11 is induced only after a medium shift into ncPro medium.

Our first efforts¹⁹ utilized split GFP1-10/11 (Figure 5.2b), a pair of protein fragments initially used to screen for protein solubility.²⁰ With this approach, the proline-containing GFP fragment GFP1-10 (containing the first ten β strands of GFP) was constitutively expressed with a modified version of the strong *tac* promoter²¹ (missing the LacI binding site) on a pBAD33 backbone. Two proline-free GFP11 strands, separated by an elastin-like

protein linker, were placed under control of the IPTG-inducible T5 promoter on a pQE-80L backbone. One proline codon preceded the first GFP11 domain. With the five proline codons in the elastin linker, expression of the full GFP11-elastin-GFP11 fusion protein (which will henceforth generally be referred to as “GFP11”) was dependent upon readthrough of six proline codons. Complementation of the two GFP fragments, an association reported to occur with sub-picomolar affinity,²² should form the completed β -barrel of GFP. Subsequent chromophore formation would provide a fluorescent readout of ncPro incorporation. The scheme outlining initial expression conditions is shown in Figure 5.2c. Finally, *proS*, the gene encoding the *E. coli* ProRS, was placed under control of its endogenous promoter and on the same plasmid backbone as GFP11 (Figure 5.2b). We envisioned replacing this wild-type ProRS with a library of mutants when screening for improved variants.

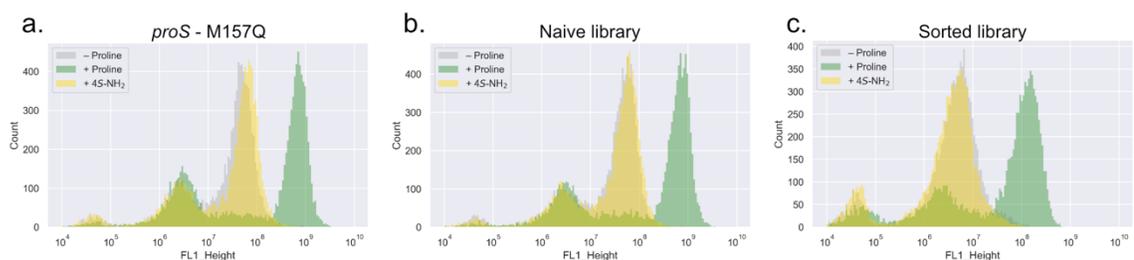


Figure 5.3. 4S-NH₂ incorporation does not improve after sorting. Flow cytometry after a split-GFP1-10/11 expression experiment for *E. coli* strain SLB2001 overexpressing the following ProRS variants or libraries: **a.** ProRS-M157Q; **b.** Error-prone PCR library. **c.** 1x sorted library. The fluorescence of the population treated with 4S-NH₂ did not increase relative to the negative (–Proline) control after sorting.

Initial efforts appeared promising: separation could be discerned between +/- proline samples when using *E. coli* strain SLB2001, a proline auxotroph derived from DH10B. Further, low levels of 4S-aminoproline (4S-NH₂) incorporation could be observed as a

slight increase in fluorescence over background (Figure 5.3a) when overexpressing the M157Q ProRS mutant, which is known to modestly improve 4S-NH₂ incorporation.¹⁹ However, sorting an error-prone PCR library of ProRS variants for incorporation of 4S-NH₂ did not improve relative fluorescence (Figure 5.3b-c).

5.4.3 Efforts to improve split-GFP dynamic range reveals poor GFP1-10 solubility

Several control experiments gave rise to concern. First, we noted high background fluorescence without adding proline or a ncPro after the medium shift, or inducing GFP11 expression (Figure 5.4a). We did not detect an increase in fluorescence over background for strains containing the GFP1-10 gene alone (the –GFP11 population is representative of *E. coli* autofluorescence only; data not shown), suggesting that these results are due to leaky GFP11 expression, and proline codon read-through. We did not detect incorporation of any other canonical amino acid into GFP11 in the absence of proline by mass spectrometry (Figure 5.4b-c). Increasing wash step stringency during the medium shift did not have any effect on background fluorescence (data not shown).

Residual proline incorporation appears to be a general phenomenon in all proline replacement approaches described in this thesis. Low levels of proline-containing proinsulin are nearly always observed by mass spectrometry, even when using proline analogs with high incorporation efficiencies (Figure 2.2a-d, Figure 3.2a-d), and low levels of proinsulin can be observed by SDS-PAGE even in the absence of proline (Figure 2.S1).

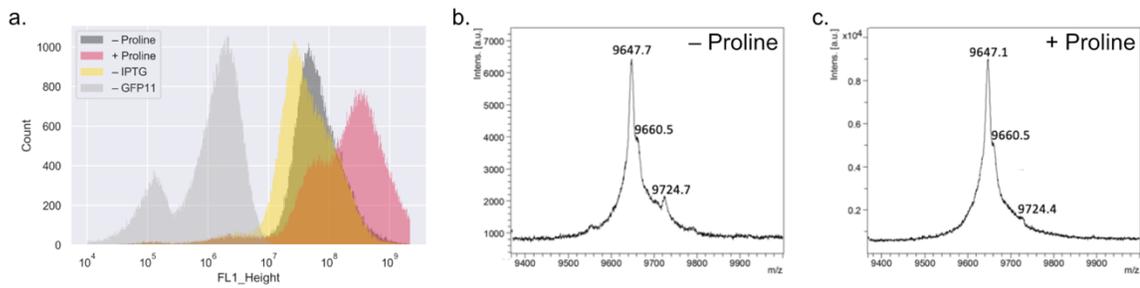


Figure 5.4. Leaky GFP11 expression and residual proline result in high background fluorescence. **a.** Fluorescence of the split-GFP system as measured by flow cytometry. Negative controls include a strain lacking the GFP11 gene (–GFP11), no induction of GFP11 expression (–IPTG), and no addition of proline (–Proline). **b-c.** MALDI-TOF spectra of purified GFP11 expressed in the absence (b) or presence (c) of proline. Expected m/z of proline-containing GFP11 fragment = 9647.5. No significant levels of amino acid misincorporation were detected.

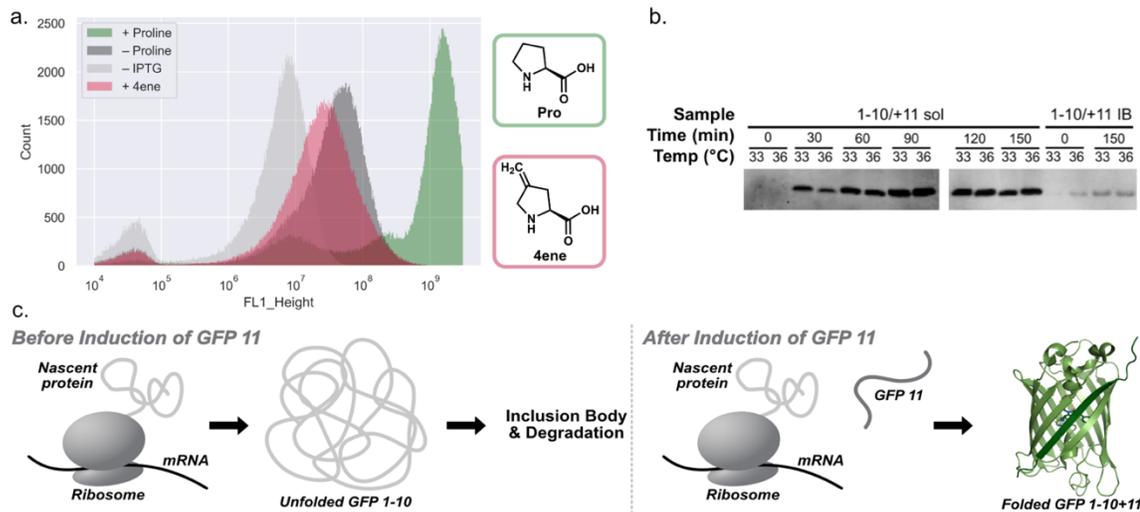


Figure 5.5. Poor solubility impairs split-GFP performance. **a.** Split-GFP flow cytometry results. 4-ene (~90% incorporation efficiency, Chapter II) is less fluorescent than the “–Proline” control. **b.** Western blot detecting the presence of GFP1-10 in the soluble (sol) or inclusion body (IB) fraction over time after inducing GFP11 expression; similar effects were observed at both 33 and 36 °C. **c.** Proposed model for GFP1-10/11 expression in *E. coli*. In the absence of GFP11, GFP1-10 is insoluble and unstable in *E. coli*. In the presence of GFP11, protein stability and chromophore formation are achieved.

Tighter control of GFP11 expression with the *araBAD* promoter²³ moderately improved the resolution between positive and negative controls (+/- Proline, Figure 5.5a). However, treatment with the proline analog 4ene, which exhibits high levels of incorporation during proinsulin expression (see Chapter II of this thesis), did not result in a fluorescent population commensurate with the expected incorporation efficiency. In fact, this population was *darker* than the negative control, a sample which did not contain any additional proline (Figure 5.5a). This result was reminiscent of the mWasabi example above (Figure 5.1b).

A western blot analysis of the soluble and insoluble fractions illustrated a significant difference in protein solubility in the presence and absence of GFP11 (Figure 5.5b). Before induction of GFP11 expression (time 0), no GFP1-10 was detected by western blot at either temperature tested. However, robust expression could be detected in the soluble fraction 30 minutes after inducing GFP11 expression, and grew more prominent over time. Faint bands could be detected in the insoluble, inclusion body fraction before and after GFP11 expression.

We propose that in the absence of GFP11, GFP1-10 does not fold well, and either is actively degraded by the cell, or is insoluble. As a result, proline-containing GFP1-10 expressed before the medium shift is unavailable for GFP11 complementation after ncPro incorporation. In the presence of GFP11, newly synthesized GFP1-10 remains soluble by associating with GFP11, and is capable of chromophore formation (Figure 5.5c). Importantly, the only GFP1-10 able to associate with GFP11 is that translated *after* the medium shift and *in the presence of the ncPro*, so the advantage of using split-GFP over full-length fluorescent proteins is lost. While many others have used this particular split-

GFP pair in a wide range of contexts,²⁴⁻³¹ we are not aware of any studies that require temporal separation of each GFP fragment. In fact, in manuscripts describing uses of GFP1-10/11, GFP1-10 is either expressed after the GFP11 fragment, or is purified from the inclusion body fraction and then refolded.²⁵ We have also observed that strains constitutively expressing GFP1-10 under control of the modified *tac* promoter exhibit substantially slower growth rates (data not shown), suggestive of GFP1-10 toxicity.

5.4.4 Solubility tags improve GFP1-10 solubility

Decreasing expression temperature to 25°C did restore some GFP1-10 solubility in the absence of GFP11, albeit at reduced levels (Figure 5.6a). To better control protein expression, the constitutive *tac* promoter previously used to drive GFP1-10 expression was replaced with the IPTG-inducible *T5* promoter (Figure 5.6b). We envisioned inducing GFP1-10 expression in proline-containing medium at mid-log phase. After incubation for 3 h at 25°C, cells would be shifted into a proline-free medium. The ncPro of interest was added under osmotic stress conditions, and expression of the GFP11 fragment induced (Figure 5.6c). However, no significant increase in fluorescence could be detected with the addition of proline (Figure 5.6d). These data suggest that, despite the apparent increased solubility at lower temperatures, there are still insufficient quantities of soluble GFP1-10 available for complementation after the medium shift.

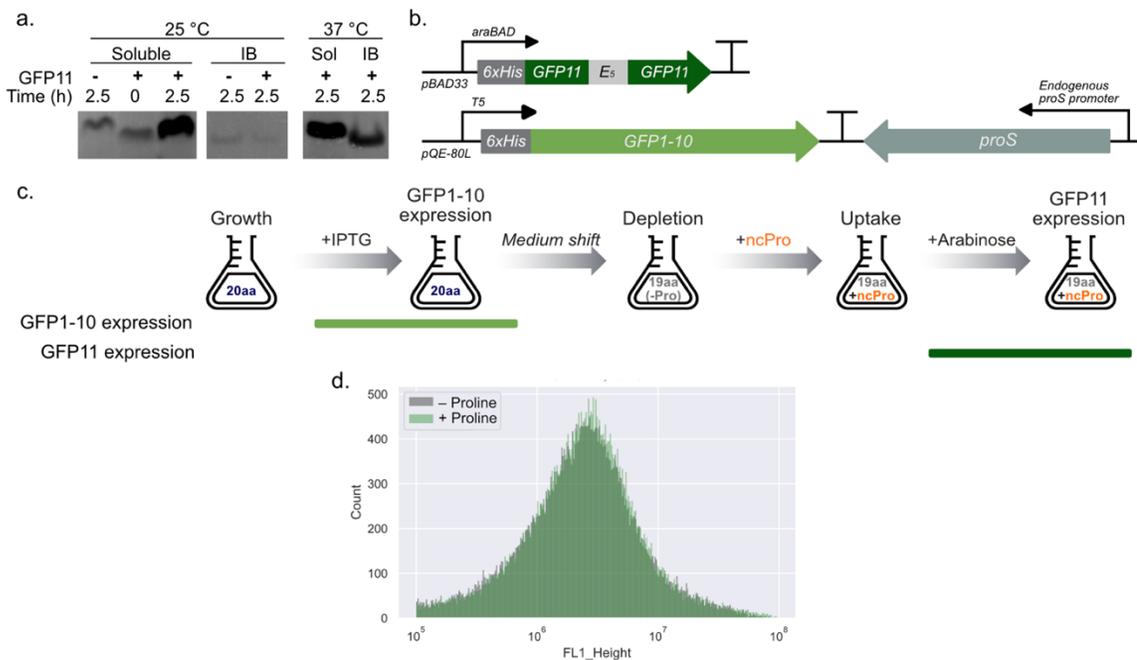


Figure 5.6. Staggered expression of split-GFP at 25°C. **a.** Western blot analysis of GFP1-10 in the soluble and inclusion body (IB) fractions during split-GFP expression. GFP1-10 was expressed constitutively, and expression of GFP11 was induced at time=0 at 25 and 37°C. **b.** Redesigned plasmid scheme that includes inducible control of expression for both GFP fragments. **c.** Staggered expression protocol. Expression of GFP1-10 is induced at mid-log phase in proline-containing medium prior to a medium shift to proline-free medium. GFP11 expression is then induced in the presence of the ncPro. Association and fluorescence are dependent upon the ability of soluble GFP1-10 to persist throughout the experiment. **d.** Fluorescence in the presence and absence of proline, as measured by flow cytometry.

“Supercharging” proteins is an approach to solubilizing proteins that involves replacing nonpolar surface residues with charged amino acids.³² Both super-positive and super-negative versions of GFP with net charges of +36 and -30, respectively, have been developed. We cloned the corresponding split super-positive and -negative versions of GFP1-10 (referred to as “spGFP” and “snGFP”) and assessed their ability to complement GFP11. However, even simultaneous expression of both GFP1-10/11 fragments did not result in an increase in fluorescence, as measured by plate reader (Figure 5.7a,b). These

data indicate that chromophore formation is inhibited for these charged GFP variants, perhaps due to decreased fragment association.

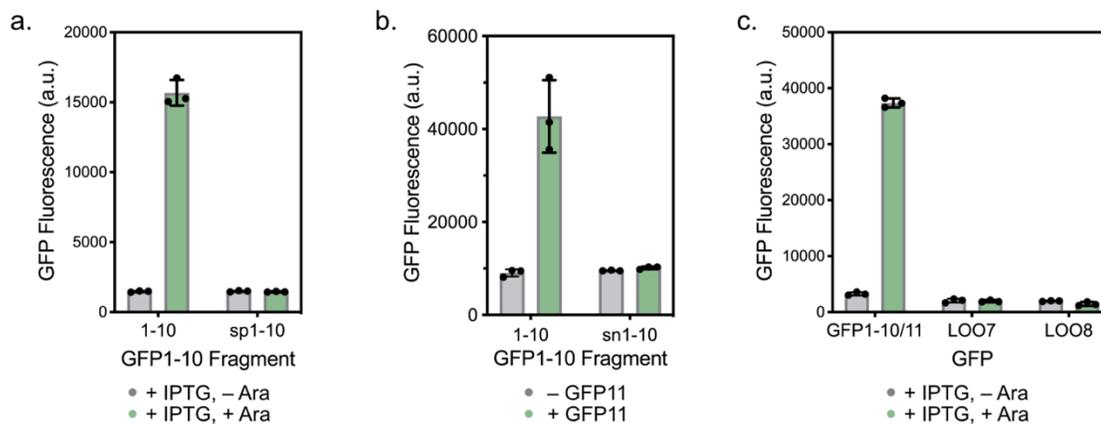


Figure 5.7. Fluorescence of split supercharged and permuted GFP variants. IPTG and arabinose induce expression of the 1-10 and 11 fragments, respectively. Complementation was assessed by co-expressing each GFP1-10/11 fragment. **a.** sp1-10/GFP11 complementation was assessed in rich medium after inducing expression of each GFP fragment. **b.** sn1-10/GFP11 complementation was assessed in rich medium by measuring fluorescence of strains containing or lacking the GFP11 gene. **c.** Complementation of circularly permuted GFP fragments was assessed in minimal medium after inducing expression of each GFP fragment.

“LOO7” and “LOO8” are versions of split-GFP in which either the 7th or 8th strand of the β -barrel of a circularly permuted GFP variant is omitted.³³ Neither the 7th or 8th strand of GFP contains proline residues, and the resulting large fragment was reported to be either more soluble (LOO7) or lead to greater fluorescence (LOO8) than the split-GFP1-10/11 version.³³ However, we did not observe fluorescence upon co-expression of both fragments for either split-GFP variant tested (Figure 5.7c).

A common approach to increase protein solubility is fusion to a solubility tag. We translationally fused MBP, TrxA, SUMO, or NusA to the N-terminus of the GFP1-10 fragment. In initial experiments, the constitutive *tac* promoter described above was used to

drive expression of the GFP1-10 fragments; GFP11 expression was controlled by the T5 promoter (Figure 5.2b). We probed the solubility of each GFP1-10 fusion protein by western blot in the presence and absence of GFP11. All fusion proteins were found in the soluble fraction when expressed in the presence of GFP11. In the absence of GFP11, small but detectable levels of 1-10 fusion proteins could be observed in the soluble fraction (Figure 5.8a).

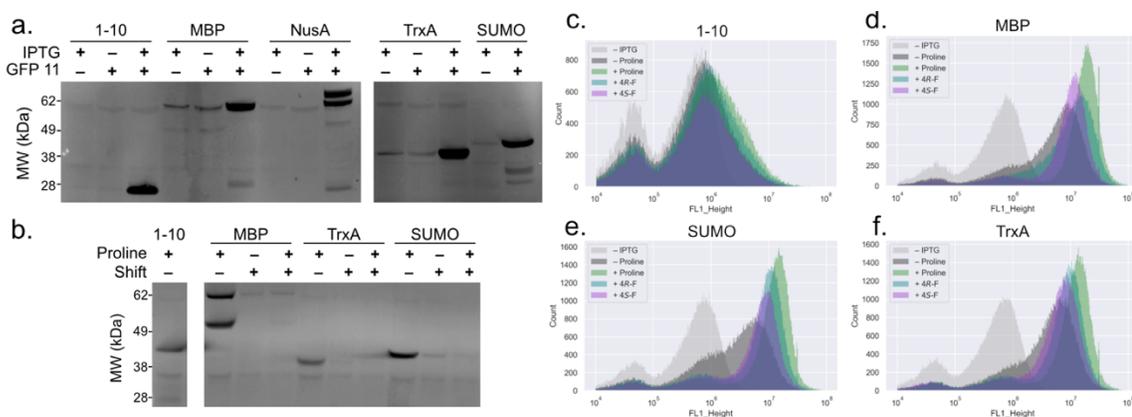


Figure 5.8. GFP1-10 solubility tag fusions improve resolution by flow cytometry. **a.** Western blot analysis of constitutively expressed GFP1-10 fragments in the soluble fraction at 37°C in the presence or absence of the GFP11 gene, and with or without inducing GFP11 expression with IPTG. **b.** Western blot analysis of GFP1-10 in the soluble fraction during expression in minimal medium at 25°C. Samples were collected either before or after the medium shift; GFP1-10 without a solubility tag was expressed in rich (LB) medium. **c-f.** Flow cytometry analysis of GFP1-10, or GFP1-10 fusion proteins. Flow cytometry results were obtained from staggered expression of each GFP1-10/11 fragment: GFP1-10 was first expressed in proline-containing medium, and GFP11 was expressed after a medium shift.

To measure protein solubility under conditions that more closely mimic ncPro incorporation, we grew cells in M9 medium, performed a medium shift, and controlled expression of both split-GFP fragments with the pBAD/pQE80 construct design (Figure 5.6b). We observed soluble GFP1-10 at 25°C even in the absence of GFP11 (Figure 5.8b),

though levels of each GFP1-10 fusion protein in the soluble fraction decreased substantially after a medium shift.

Tagged split-GFP constructs were assessed by flow cytometry. In the absence of a solubility tag, minimal levels of fluorescence over background were observed (Figure 5.8c), consistent with poor GFP1-10 solubility even at lowered temperatures. For the tagged GFP1-10 constructs, we did observe an increase in fluorescence for the +Proline positive control. Further, samples treated with 4*R*-F and 4*S*-F (proline analogs with distinct conformational behaviors⁵) both exhibited fluorescence (Figure 5.8d-f), suggesting that this may be a general approach to assess ncPro incorporation. However, the resolution between the positive (+Proline) and negative (–Proline) controls was low, and likely insufficient for effective library sorting by FACS.

5.4.5 Changing expression protocols does not improve screen performance

Based on the observation that the medium shift reduces the levels of GFP1-10 in the soluble fraction (Figure 5.8b), we hypothesized that performing a medium shift increases protein degradation in *E. coli* and leads to reduced dynamic range. We explored alternative GFP1-10/11 expression protocols, in which the expression of GFP1-10 would be induced *after* the expression of GFP11 (Figure 5.9a-b). This protocol would rely upon the stability of GFP11, rather than GFP1-10. However, these approaches did not improve screen performance (Figure 5.9c-d).

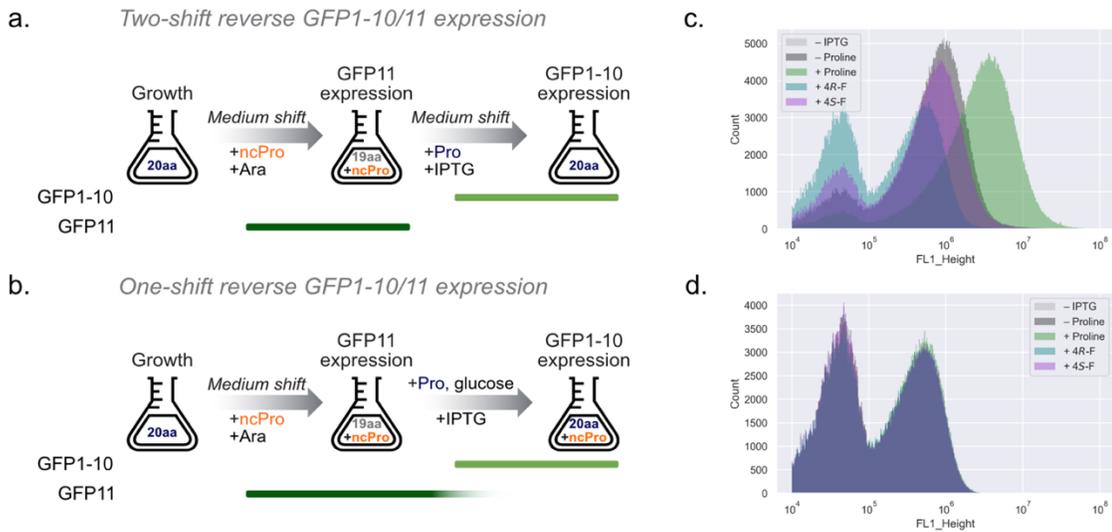


Figure 5.9. GFP1-10/11 reverse expression. **a.** Expression protocol for a two-shift reverse GFP1-10/11 expression protocol. GFP11 expression is induced after the first shift to ncPro medium; GFP1-10 expression induced after a second shift to proline medium that contained glucose as a carbon source to repress GFP11 expression. **b.** One-shift reverse protocol: proline (to out-compete the ncPro) and glucose (to repress GFP11 expression) were added before inducing expression of GFP1-10. **c-d.** Flow cytometry results for two-shift (**c**) and one-shift (**d**) ncPro incorporation experiments.

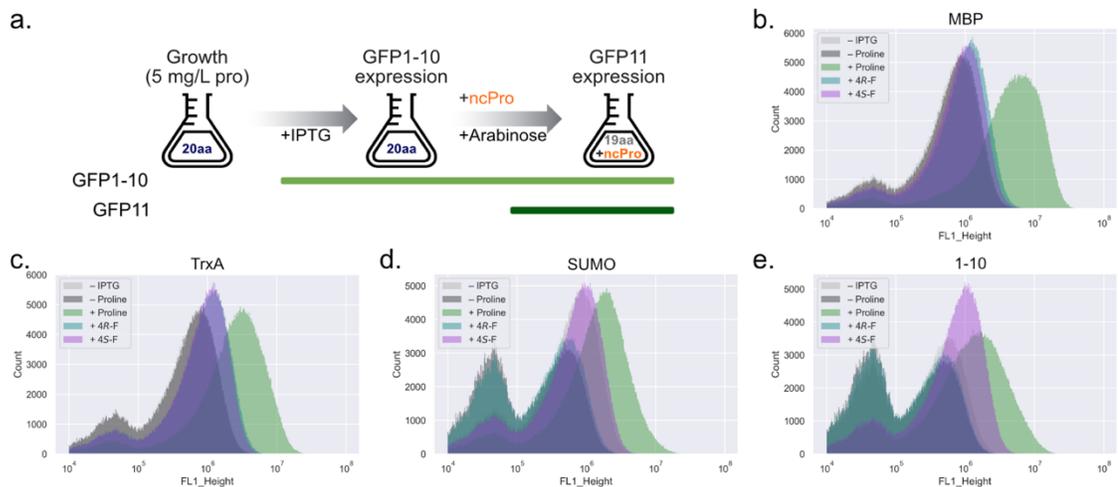


Figure 5.10. SPI expression of GFP1-10/11. **a.** Expression protocol for SPI-based expression of split-GFP. **b-e.** Flow cytometry results for ncPro incorporation experiments with MBP-GFP1-10 (**b**) TrxA-GFP1-10 (**c**), SUMO-GFP1-10 (**d**), or GFP1-10 (**e**).

An alternative method used for residue-specific incorporation of ncAAs is selective pressure incorporation (SPI).³⁴ Rather than use a medium shift, this approach relies upon bacterial growth in amino acid limiting medium, before adding the ncAA. In the context of GFP1-10/11 expression (Figure 5.10a), SPI did improve resolution between the +Proline and –Proline controls by drastically reducing background in the absence of proline. However, samples treated with the translationally active 4-fluoroproline tested were barely fluorescent above background (MBP–1-10 and TrxA–1-10, Figure 5.10b-c), or did not demonstrate similar levels of fluorescence between the two diastereomers (SUMO–1-10 and GFP1-10, Figure 5.10d-e).

5.4.6 Alternative GFP11 designs do not improve performance

XTEN is an unstructured protein commonly used to increase the half-life of therapeutic proteins and peptides.³⁵ This domain contains many proline residues, and its length can be easily changed to tune screening stringency. We hypothesized that, because of its disordered, hydrophilic nature, ncPro incorporation would not affect its solubility. We constructed two versions of XTEN-GFP11 fusion proteins: XTEN72-GFP11, which contains a 72-residue XTEN domain with 13 proline codons N-terminally fused to the GFP11 fragment; and XTEN144-GFP11, which contains 25 residues N-terminal to GFP11. These genes were installed in the pBAD33 plasmid, replacing the GFP11-elastin-GFP11 gene shown in Figure 5.6b. We expressed the MBP–GFP1-10 fusion protein for 3 h at 25°C. After a medium shift and addition of ncPro residues, XTEN-GFP11 expression was induced with arabinose. We measured the fluorescence of each construct after 135 min of GFP11 fragment expression, and compared it to the original GFP11-elastin-GFP11 design (Figure 5.11).

As expected, we find that including more proline codons upstream of the GFP11 gene does improve the resolution between the –Proline and +Proline controls. The 25 proline codons in XTEN144 are too stringent for ProRS evolution efforts: even 4*R*-F, which is incorporated with high efficiency,⁵ is barely discernible over background for this strain. Fluorescence of the 4*S*-F treated XTEN samples is most concerning, however. Compared to the original GFP11-*elastin*-GFP11 design, in which both 4*R*-F and 4*S*-F display fluorescence levels between the positive and negative controls (Figure 5.11a), the 4*S*-F–treated XTEN samples are not fluorescent over background (Figure 5.11b-c). These results suggest that 4*S*-F incorporation is not well tolerated by XTEN, perhaps due to low protein solubility after incorporation, and demonstrate that this design is not a good general approach to measure ncPro incorporation.

Other attempts at improving resolution of our split-GFP system using alternative proline auxotrophic strains of *E. coli*, or knocking out the *lon* protease, were similarly unsuccessful (data not shown).

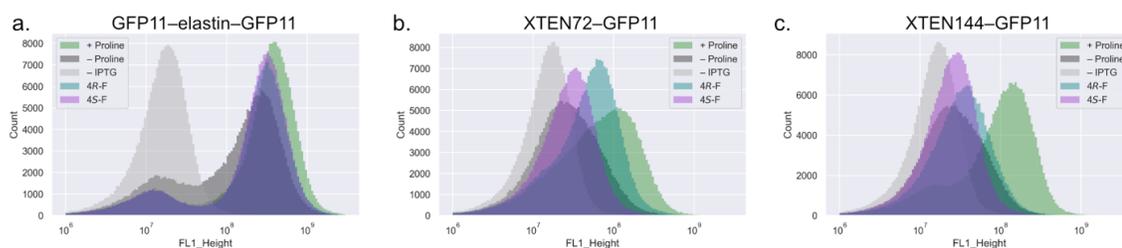


Figure 5.11. XTEN-GFP11 fusion proteins to measure proline analog incorporation. Cells were grown in minimal medium until early log phase, at which point expression of the MBP–GFP11-10 fusion protein was induced for 3 h at 25°C. After a medium shift and ncPro addition, the original GFP11 fragment (a), or GFP fused to the C-terminus of XTEN72 (b) or XTEN144 (c), was expressed for 135 min. Fluorescence was assessed by flow cytometry.

5.4.7 Inclusion body analysis by flow cytometry

As detailed earlier in this thesis, we are interested in applying ncPro mutagenesis to study and modulate the behavior of the peptide therapeutic insulin. To prepare our insulin variants, we first express proinsulin, a precursor to insulin, in *E. coli*. Proinsulin is isolated from the inclusion body fraction before refolding and maturation. Inclusion body formation in *E. coli* has been monitored by measuring the scattering properties via flow cytometry;³⁶ we wondered if a similar technique could be used to assess proinsulin production, which correlates well with ncPro incorporation efficiency (Figure 2.S1, Table 2.S1). While the resolution and dynamic range of such an approach is expected to be low, it would have the advantage of directly observing proinsulin production (rather than monitoring the expression of a reporter protein). To assess feasibility, we measured the scattering properties of *E. coli* by flow cytometry after a typical proinsulin expression experiment. However, no difference in scattering was observed after inducing proinsulin expression (Figure 5.12).

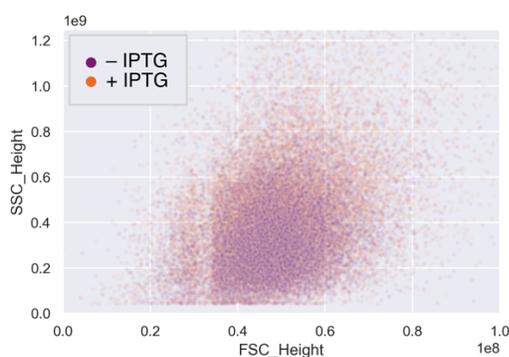


Figure 5.12. Interrogating inclusion body formation by flow cytometry. Forward-scatter (FSC) vs side-scatter (SSC) properties of *E. coli* with (+IPTG) or without (-IPTG) inducing proinsulin expression, as measured by flow cytometry.

5.4.8 *Survival on proline analog-containing medium*

Because our ncPro incorporation strategy relies upon a residue-specific approach that leads to proteome-wide proline replacement, we hypothesized that extended ncPro exposure might lead to toxicity. Differential growth between high- and low-ncPro incorporating strains, determined after replica plating, might be used to screen for improved ncPro incorporation. Here, inhibited growth in the presence of the ncPro would indicate higher levels of incorporation.

Strains overexpressing either the wild-type ProRS, or M157Q mutant (which increased incorporation of the ncPro 4S-NH₂, Ref. 19) were plated onto agar plates with rich (LB) medium. We transferred colonies by replica plating to M9 medium agar plates that contained low (1-10 μ M) concentrations of proline, and increasing (0-500 μ M) concentrations of 4S-NH₂. However, we did not observe differences in colony formation as a function of known 4S-HN₂ incorporation efficiency (data not shown).

5.4.9 *A small molecule alternative to fluorescent reporter proteins*

Small molecule organic dyes possess several advantageous properties compared to fluorescent proteins, such as increased brightness and photostability.³⁷ Since a small molecule dye is not itself genetically encoded, but can be designed to specifically label a protein of interest, judicious choice of a labeling target might avoid the ncPro-specific effects on full-length fluorescent proteins.

FlAsH-EDT₂ is a small organoarsenic molecule derived from the fluorophore fluorescein (Figure 5.13a). In its unbound state, FlAsH-EDT₂ is not fluorescent. However, upon binding to a tetra-cysteine (TC) peptide motif, increased conformational restriction leads

to fluorescence upon irradiation.³⁸ FIAsh is especially useful for protein localization studies: compared to fusing a protein of interest to a large fluorescent protein, which can lead to artifacts in imaging and protein behavior,³⁹ the FIAsh system requires only a short peptide sequence.³⁸ Despite these advantages, FIAsh is not widely used, perhaps due to high background labeling.⁴⁰

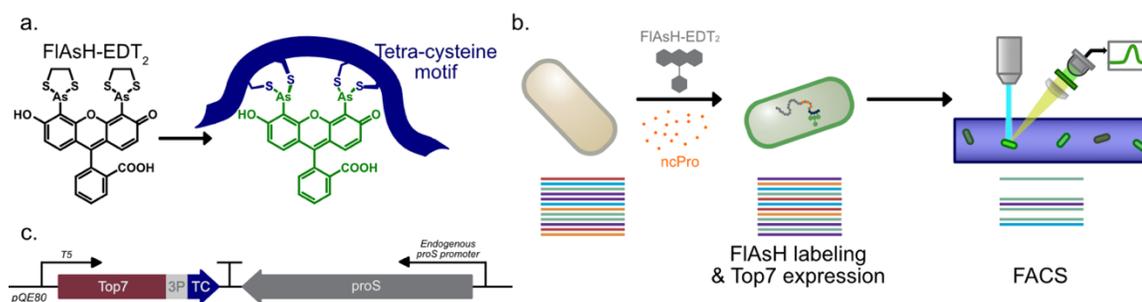


Figure 5.13. A FIAsh-based labeling strategy to measure proline analog incorporation. **a.** The chemical structure of the biarsenical fluoresceine derivative FIAsh-EDT₂. FIAsh fluoresces after binding to a tetra-cysteine (TC) motif, which can be translationally fused to a protein of interest. **b.** Proposed protocol for ProRS engineering by FIAsh labeling. A library of ProRS variants (represented by the colored lines) are expressed in *E. coli*. A TC motif is translationally fused to a soluble, proline-free protein and proline-containing linker. FIAsh labeling detects TC motif expression after ncPro incorporation, which can be detected by flow cytometry. FACS would sort cells based on fluorescence, leading to enrichment of ProRS library members capable of enhanced ncPro incorporation. **c.** Plasmid design. The expression of Top7-TC3 is controlled by the IPTG-inducible T5 promoter; the ProRS library is installed on the same pQE-80L plasmid backbone.

We envisioned applying a FIAsh-based labeling approach to measure ncPro incorporation, and took several design considerations into account. First, we identified Top7 as a candidate carrier protein to which the TC motif could be fused. Top7 is the result of a computational protein design endeavor, and, importantly, contains no proline residues.⁴¹ Its behavior (i.e., solubility or susceptibility to degradation) should not be dependent upon

the ncPro residue of interest. Second, we appended a proline-containing linker sequence to the C-terminus of Top7 that would connect the carrier protein to the TC motif. Readthrough of the proline codons in this linker, presumably by ncPro incorporation, would allow for translation to continue to the TC motif. Finally, we considered the TC motif itself. The generic TC motif is the amino acid sequence CCXXCC, where X represents any amino acid. However, an internal proline-glycine dipeptide (as in the TC motif CCPGCC) has been found to most effectively engage FAsH, since the fluorophore binds best to β -hairpin turns.⁴² Because fluorescence after FAsH binding to a CCPGCC TC motif depends on the ncPro incorporated (Figure 5.S1), we elected to search for alternate TC motifs that do not contain proline residues. Sorting a library of proline-free TC motifs by FACS (Figure 5.S2) identified TC3, whose amino acid sequence is YCCGVCCI.

Typical FAsH labeling experiments proceeded as follows (Figure 5.13b). An *E. coli* proline auxotrophic strain harboring plasmid pQE80_Top7-TC3_proS, which contains an IPTG-inducible Top7-TC3 gene and the *E. coli proS* under the control of its endogenous promoter (Figure 5.13c), was grown at 37°C to mid-log phase in minimal medium that contained proline. Cells were washed with cold 0.9% NaCl, and resuspended in medium lacking proline. FAsH was added at this stage, since labeling cells after Top7 expression led to poor fluorophore uptake and significant heterogeneity (Figure 5.S3). After 30 min incubation to deplete residual proline, the ncPro of interest was added to the culture. High (0.3 M) levels of sodium chloride were also added to facilitate ncPro uptake.⁶ After inducing expression of Top7-TC3 with IPTG, samples were washed and analyzed by flow cytometry; fluorescence should correlate with ncPro incorporation.

5.4.10 Optimization of FIAsh labeling conditions does not improve sorting performance

We tested a range of proline auxotroph strains, and found that all yielded similar results by flow cytometry (Figure 5.14a-e). We chose to continue with strain CAG18515, since we use this strain for proinsulin expression. The bright tail present in many samples was removed through more stringent wash steps that use 3% DMSO and ethanedithiol (EDT); representative histograms of Top7 expression and optimized FIAsh labeling and wash conditions are shown in Figure 5.14f.

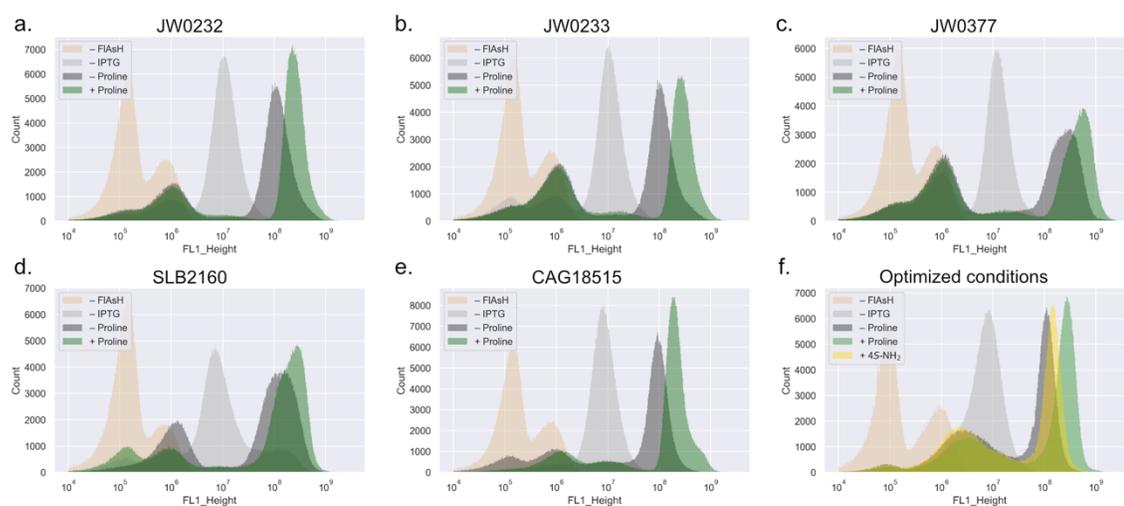


Figure 5.14. Optimization of FIAsh labeling and protein expression conditions. a-e. Flow cytometry histograms of proline auxotrophic strains of *E. coli* after Top7-TC3 expression and non-optimized FIAsh labeling. Strain SLB2160 is ampicillin resistant, so plasmid pQE80-c_Top7-TC3_proS, which confers resistance to chloramphenicol, was used in that case. f. Lower concentrations of FIAsh added during labeling, and more stringent washes, improves fluorescence histogram appearance; shown are optimized labeling conditions for strain CAG18515.

We cloned a site-saturation mutagenesis library by standard restriction enzyme cloning approaches, targeting five residues in the ProRS active site (Figure 5.15a). Extreme care was taken to purify correctly-sized plasmid DNA by agarose gel. Nevertheless, sorting the

brightest ~0.1% of cells after treatment with 4S-NH₂ did not result in increased fluorescence, and instead led to a more heterogeneous population (Figure 5.15b-c). Preliminary analysis of the sorted plasmid DNA suggested loss of the *proS* gene (data not shown). We also noted low rescue efficiencies in our sorting experiments: after rescuing sorted cells in rich medium at 37°C, <1% of our sorted events led to colony-forming units (CFUs) when plated onto selective agar plates.

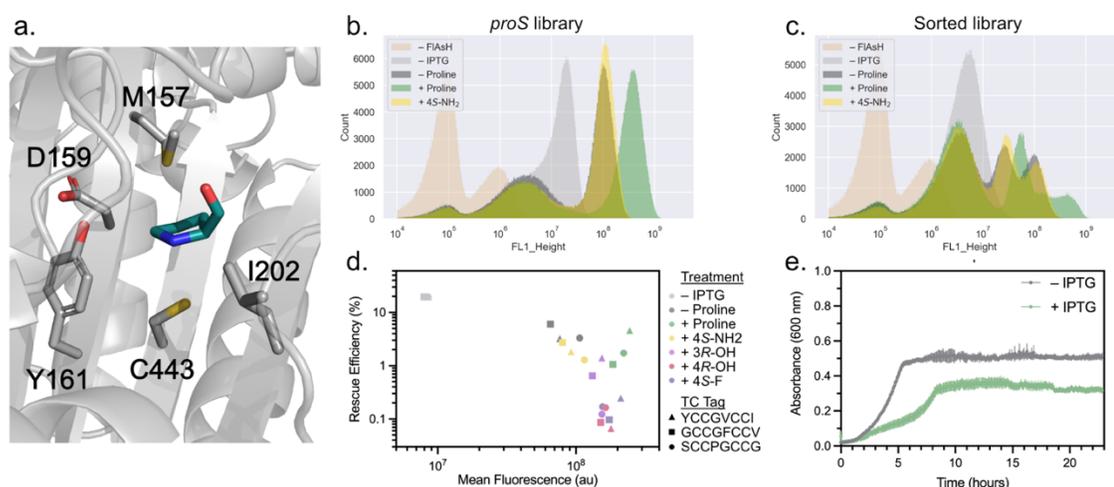


Figure 5.15. Sorting after FIAsh labeling reveals poor rescue efficiency and Top7 toxicity. a. Homology model of the *EcProRS* bound to prolinol (teal), an analog of proline used in the initial crystal structure. The model was created using SWISS-MODEL⁴³ using the structure of the ProRS from *Enterococcus faecalis* (PDB: 2J3M, 42% homology) as the template. **b.** Flow cytometry histograms after Top7-TC3 expression and FIAsh labeling. The top ~0.1% of the sample treated with 4S-NH₂ was sorted, rescued in rich medium, and was subjected to a second round of analysis by flow cytometry (**c.**). **d.** The entire bright population of a clonal sample overexpressing the wild-type ProRS was sorted after Top7 expression and FIAsh labeling. Sorted cells were rescued in rich medium for 1 h at 37°C, then plated onto selective agar plates. Rescue efficiency was defined as CFUs divided by total events sorted. **e.** Growth with (+IPTG) and without (–IPTG) inducing Top7-TC3 expression. Cells were diluted into M9 medium with 0.3 M NaCl (to simulate ncPro incorporation conditions), and growth was monitored in a 96-well plate at 37°C with shaking.

We sorted the entire bright population (FL1 > $\sim 10^7$) under a variety of expression conditions, and found that rescue efficiency was inversely correlated with fluorescence across three TC tags assessed (Figure 5.15d). Especially notable is the decreased rescue efficiency between –Proline (which should result in *lower* viability for this proline auxotrophic strain of *E. coli*), and +Proline conditions. Further, Top7 expression substantially slowed growth (Figure 5.15e). We hypothesized that increased Top7 expression (which should indicate higher levels of ncPro incorporation) hindered growth and survival after sorting, an undesired property for a protein engineering endeavor. These effects might enable out-competition by any cells containing undesired plasmids; for instance, those found here lacking the *proS* gene.

We identified eight other soluble, proline-free proteins, tested their toxicity when expressed in *E. coli* (Figure 5.16a-h), and analyzed select strains by flow cytometry after FAsH labeling (Figure 5.16i-l). The protein DHR14, a highly stable, computationally-designed helical repeat protein⁴⁴ performed best. We note here that sample heterogeneity (in particular, the darkest population for each FAsH-labeled sample) was present in all strains assessed by flow cytometry (Figure 5.16i-l). We find that dark cells were physiologically distinct: their scattering properties differed from brighter cells (Figure 16m), and appeared elongated by confocal microscopy (Figure 16n-p). Shown are images for strains expressing DHR14-TC3; Top7-TC3 strains behaved similarly (data not shown).

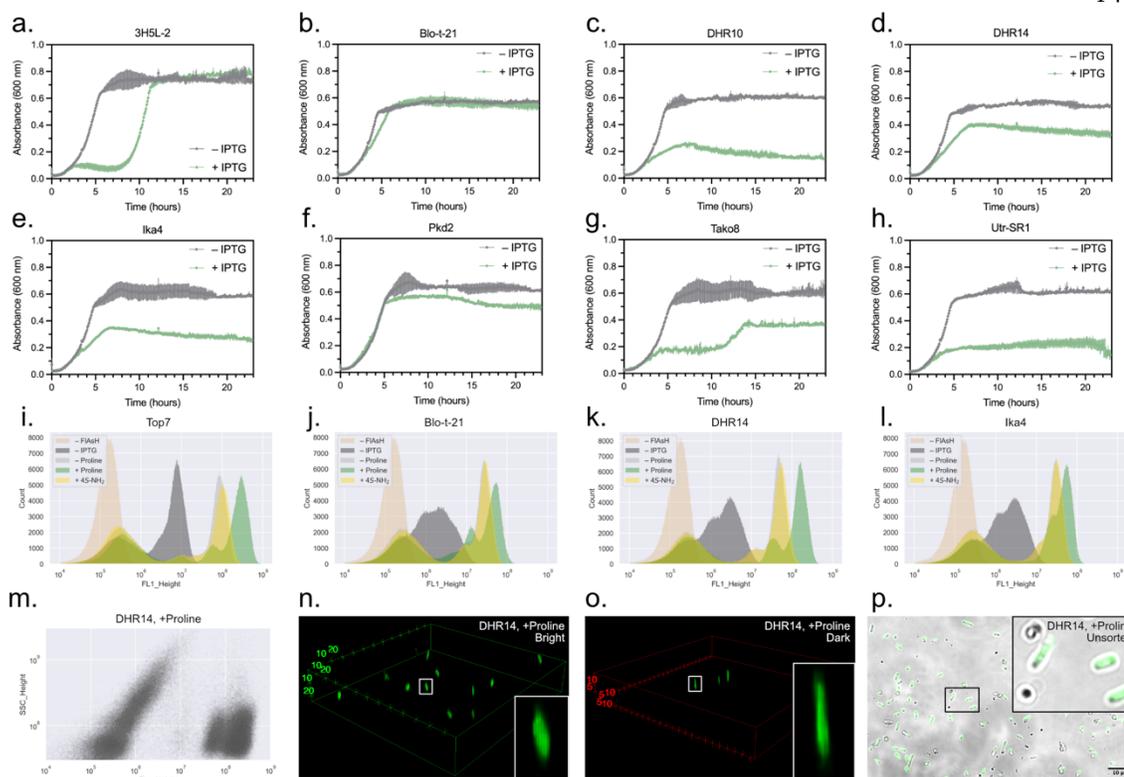


Figure 5.16. Alternative proline-free target proteins. **a-h.** Eight alternative proline-free proteins were identified, and replaced the Top7 gene in plasmid pQE80_Top7-TC3_proS. Growth of strain CAG18515 carrying each of these plasmids in M9 medium +0.3 M NaCl was monitored by tracking absorbance at 600 nm in a plate reader at 37°C with shaking. **i-l.** Flow cytometry histograms of select proline-free proteins after ncPro incorporation and FIAsh labeling. We chose carrier proteins whose expression limited growth the least; Pkd2 was omitted because we could not detect its expression by SDS-PAGE (data not shown). Note that the shoulder of the brightest population is likely associated with oxidized EDT: its presence disappears with fresh EDT (e.g., Fig. 5.14f). **m.** Fluorescence (FL1) versus side-scatter (SSC) of proline-treated DHR14-TC3 in panel k; all other FIAsh-labeled samples behaved similarly. **n-p.** Confocal microscopy images after DHR14-TC expression and FIAsh labeling for proline-treated samples. Bright (n) and dark (o) populations were sorted, then imaged with different microscope settings that optimized dynamic range for each sample. Unsorted cells (p) illustrate the difference in fluorescence and morphology; here, a brightfield image is overlaid with the fluorescence z-stack image.

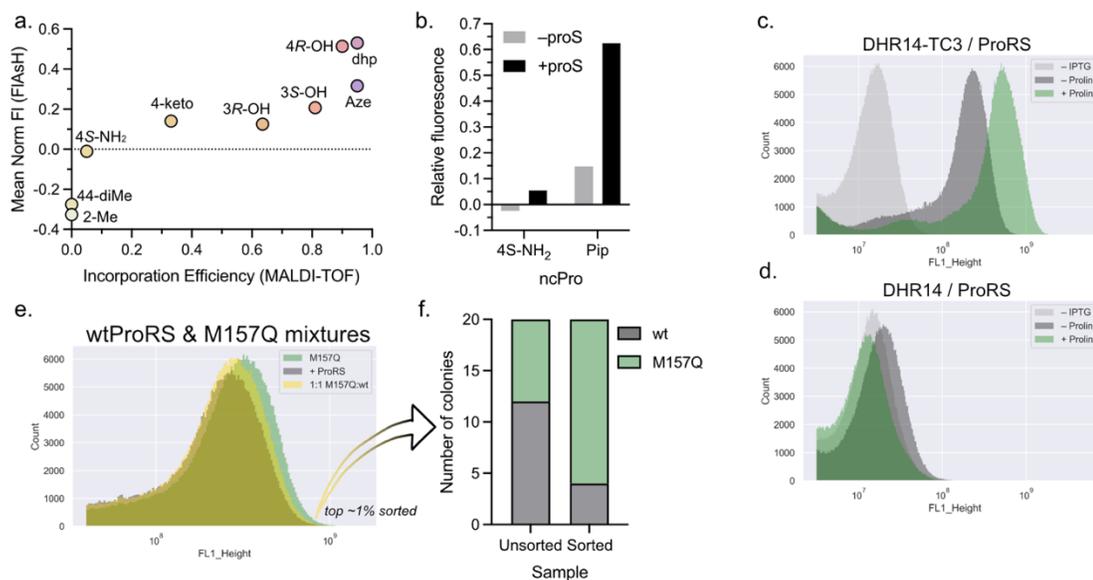


Figure 5.17. Validation of the DHR14 approach. **a.** Fluorescence, as determined by flow cytometry after DHR14-TC3 expression and FIAsH labeling, as a function of incorporation efficiency, as determined by MALDI-TOF after proinsulin expression. Flow cytometry histograms are shown in Figure 5.S4. **b.** ProRS overexpression increases fluorescence of ncPro-treated samples relative to –Proline and +Proline controls. Flow cytometry histograms are shown in Figure 5.S5. **c-d.** +Proline, –Proline, and –IPTG controls for DHR14 constructs translationally fused to TC3 (c), or lacking a TC tag (d). **e-f.** Strains overexpressing the wild-type ProRS, or mutant M157Q (which is known to increase incorporation of 4S-NH₂, ref. 19) were mixed 1:1 after DHR14-TC3 expression and FIAsH labeling. The brightest ~1% of cells were sorted (e). Plasmid DNA was purified and used to transform fresh stocks of strain CAG18515. Individual colonies were cultured, and plasmid DNA was sequenced to assess the presence of the M157Q mutation (f).

Nevertheless, we were hopeful that reduced toxicity associated with DHR14 might facilitate sorting by reducing the selective pressure against bright cells. We were encouraged by the observation that fluorescence correlated well with known ncPro incorporation efficiencies ($R^2 = 0.834$; Figure 5.17a, Figure 5.S4). Further, we observed decreased fluorescence for ncPro samples in the absence of ProRS overexpression (Figure 5.17b, Figure 5.S5), and no fluorescence over background was detected in the absence of the TC3 tag (Figure 5.17c-d). In an effort to reduce differences in physiology and rescue

efficiency between library members, we chose to lyse cells after sorting and transform fresh CAG18515 stocks using the purified DNA^{45,46} before the next rounds of sorting and analysis. Encouragingly, we could enrich the ProRS mutant M157Q after 4S-NH₂ treatment from a 1:1 mixture with the wild-type ProRS using this protocol (Figure 5.17e-f). We constructed a separate site-saturation mutagenesis library, this time only targeting four locations (excluding position 202; 4SSM) to reduce library size and cloning burden. Separately, we cloned an error-prone PCR library (ePCR), targeting an average of 2 mutations per *proS* gene (Table 5.S1).

Despite these changes, we still observed significant increases in sample heterogeneity after sorting across the three ncPro analogs and two libraries assessed (Figure 5.18a-b, Figure 5.S6). This darker population retained resistance to ampicillin (the plasmid selection marker; Figure 5.18c), so were not the result of non-transformed cheaters. Sequencing revealed small plasmid products that were perhaps the result of DNA recombination. These undesired plasmids lacked both the DHR14-TC3 and *proS* genes. We noted the loss of the *proS* gene in the Top7-TC3 case described earlier; presence of the Top7-TC3 gene was not assessed. Though we did not perform whole-plasmid sequencing after this earlier sorting attempt, it seems likely that similar recombination events are occurring in both cases. Culturing library strains in the absence of FIAsh led to increasing plasmid heterogeneity over multiple transformations (Figure 5.18d), suggesting that the issue is not inherent to fluorophore labeling. At this stage, because of these pervasive issues with ProRS engineering, we elected to limit our efforts in insulin modification to those ncPro residues known to incorporate well with existing methods (see Chapters II & III of this thesis).

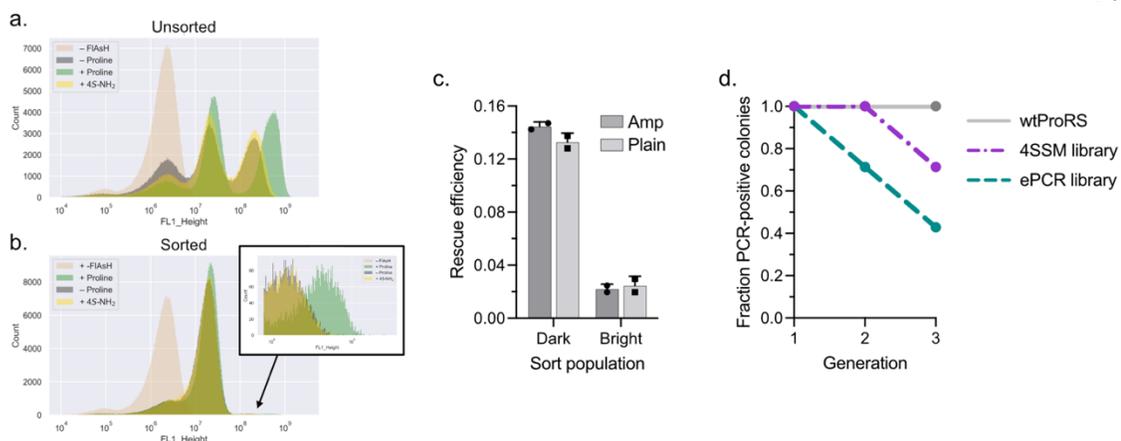


Figure 5.18. Significant problems persist. **a-b.** Flow cytometry histograms of a culture harboring a ProRS ePCR library before **(a)** and after **(b)** sorting for 4S-NH₂ incorporation. **c.** Rescue efficiencies of “Dark” ($\sim 10^7 < \text{FL1} < \sim 10^8$) and “Bright” ($\text{FL1} > \sim 10^8$) 4SSM, +Proline library populations. Sorted cells were plated onto nonselective (“Plain”) or selective (“Amp”) LB-agar plates, and rescue efficiency determined after overnight growth at 37°C. **d.** Loss of the *proS* gene over multiple generations (transformation, antibiotic selection, overnight growth, and plasmid purification). The presence of the plasmid-borne *proS* gene was assessed by colony PCR in each generation.

5.4.11 Considerations for future ProRS engineering

The most pressing issue with the FIAsh-based labeling described here is the fact that undesired plasmid species are out-competing legitimate library members. This occurs independent of recovery method after sorting (outgrowth as in Figure 5.15b-c, or plasmid purification and transformation, Figure 5.18a-b), and was found with five separate library preparations cloned by both restriction enzyme and Gibson Assembly approaches. The presence of this dark population is observed by flow cytometry across the proline-free target proteins we screened (Figure 5.16i-l), and mutagenesis seems to be occurring even in the absence of FIAsh labeling (Figure 5.18d). The exact relationship between cellular morphology (Figure 5.16n-p) and the presence of undesired plasmids is not yet clear.

The significant depletion of the desired plasmid during our sorting experiments is perplexing: we have sorted the brightest cells, yet it is this dark population that is instead consistently enriched after sorting. Mutagenesis might be occurring at a basal level throughout these experiments, in which case undesired plasmid species might continually be generated. FACS sorting is likely imperfect, which would also enable the persistence of undesired plasmids. To exacerbate the problem, strains carrying the smaller plasmids tend to grow faster, and undesired plasmids seem to lead to higher transformation efficiencies. We also note that the pQE-80L vector used here is a high copy plasmid. If DNA recombination occurs *after* target protein expression and FAsH labeling, a fraction of plasmids in a given bright cell might actually be undesired plasmid products, leading to those plasmids' persistence after sorting.

We note that our mock sorting experiment was successful in enriching a better ProRS variant after ncPro treatment (Figure 5.17e-f), and undesired plasmid species were not observed. The success of this test case compared to the failure in library sorting is puzzling. In the mock sorting case, starter cultures were grown from single colonies, whereas library cultures are necessarily heterogeneous. We did not observe plasmid instability across generations in the clonal sample (Figure 5.18d), though the relatively small number of samples and generations tested may have prevented detection if levels are initially very low. It seems likely that our library generation protocol may exacerbate the plasmid instability issue, despite the extensive quality control measures and multiple cloning approaches taken.

The strain CAG18515 may be the culprit, especially considering that mutagenesis occurred in the absence of FAsH labeling (Figure 5.18d); assessing plasmid stability in other proline

auxotrophs under these conditions might help to address the issue. Extended exposure to the organoarsenic compound FAsH, or competition molecule EDT, may also contribute to plasmid instability. For instance, inorganic arsenic is known to disrupt DNA repair pathways,⁴⁷ and could conceivably be formed as a result of FAsH metabolism in *E. coli*. FAsH likely binds to other off-target intracellular molecules (i.e., endogenous cysteine residues), which may result in unintended consequences. EDT could similarly interfere with endogenous intracellular components: for instance, reaction of EDT with carbonyl functionalities can yield 1,3-dithiolanes, and EDT is known to form metal complexes.

Our first attempts to implement this screening approach included a short FAsH labeling period after target protein expression. This protocol led to poor and very heterogeneous labeling (Figure 5.S3), a problem that was resolved by adding FAsH and EDT to cultures before inducing target protein expression (Figure 5.14f). While improving flow cytometry results from an analytical perspective, this approach led to extended incubation of cells with FAsH and EDT, which may be impacting cellular physiology and library sorting. There may be alternative labeling protocols that both limit exposure to FAsH and EDT, and result in good protein labeling.

Driving ProRS expression with the endogenous promoter might also hinder assay performance. We found that mutated plasmids generally lacked the *proS* gene, and its absence results in faster growth (Figure 5.S8a). Strains overexpressing different ProRS variants grow at different rates in M9 medium (Figure 5.S8a), which contributes an additional variable to the evolution experiment. Growing cells in rich medium (Figure 5.S8b) and controlling ProRS expression with the *araBAD* promoter (Figure 5.S8c) reduces growth differences. While we do not believe that this is the most prevalent problem

(increasing plasmid stability should be the immediate focus), we would recommend that future work drive *proS* expression with an inducible promoter.

One workaround to the presence of undesired plasmid species may be to PCR-amplify the *proS* gene from sorted cells, and use this linearized PCR product to install the sorted ProRS variants into new pQE80_DHR14-TC3_ proS vectors. A drawback to this solution is the potential accumulation of PCR biases⁴⁸ which may skew representation in a sorted library.

An alternative screening modality to FACS that avoids culturing library members together might reduce the effect of differential growth. For instance, plate reader-based assays, while lower-throughput compared to FACS, would individually culture library members in separate wells. In this case, we might be able to more easily identify high performers, even if they grow more slowly. Finally, revisiting the split-GFP system may prove more fruitful than continuing to pursue the FIASH approach. Sample heterogeneity was present by flow cytometry with this approach when expressing GFP1-10/11 in strain SLB2001 (Fig. 5.3), and this dark population was not analyzed. However, unlike performing FACS with FIASH labeling, heterogeneity was not enriched after sorting.

5.5 Conclusion

In conclusion, we have described our efforts to develop a robust and general method to screen for ncPro incorporation, and to implement it in a ProRS engineering endeavor. From an analytical perspective, our FIASH labeling approach seemed the most promising of all methods tested: we were able to achieve good correlation between known incorporation efficiencies as measured by mass spectrometry, and fluorescence as measured by flow cytometry (Figure 5.17a). Ultimately our attempts to apply this approach to ProRS

evolution were hindered by out-competition with undesired plasmid products. Together, the experiments described in this chapter lay the groundwork for future ProRS engineering endeavors. Though all approaches detailed here have yet to lead to improved ncPro incorporation, we are hopeful that future attempts at ProRS engineering can leverage this information to more successfully design and implement ProRS screening methods.

5.6 Materials and methods

5.6.1 Chemicals

All chemicals were purchased from MilliporeSigma unless otherwise indicated. FLAsH-EDT₂ was purchased from Santa Cruz Biotechnology, Carbosynth, and Toronto Research Chemicals. 1,2-ethanedithiol (EDT) was purchased from Fisher Scientific.

4-methyleneproline (4ene) was purchased as the N-boc protected version from Acros Organics, and deprotected with trifluoroacetic acid (TFA) in dichloromethane. 4ene was extracted with H₂O and lyophilized; complete deprotection and >95% purity was verified by ¹H NMR. All other proline analogs were used as received: 4*S*-aminoproline (4*S*-NH₂) was purchased as the dihydrochloride salt from Toronto Research Chemicals. 4*R*-fluoroproline (4*R*-F) and 4*S*-fluoroproline (4*S*-F) were purchased from Bachem. 2-methylproline (2-Me) was purchased as the hydrochloride salt from Advanced ChemBlocks. 4,4-dimethylproline (44-diMe) was purchased from J&W Pharmed. 4-oxoproline (4-keto) was purchased as the hydrobromide salt from Sigma Aldrich. 3*R*-hydroxyproline (3*R*-OH) and 4*R*-hydroxyproline (4*R*-OH) were purchased from Sigma Aldrich. 3*S*-hydroxyproline (3*S*-OH) was purchased from Ark Pharm. 3,4-dehydroproline

(dhp) was purchased from Combi-Blocks. Azetidine-2-carboxylic acid (Aze) was purchased from Bachem.

5.6.2 Enzymes

Restriction enzymes, kinases, and ligases were purchased from New England Biolabs (NEB). 2x Q5 master mix (NEB) was used for all PCR applications, unless otherwise indicated. DNA phosphorylation was performed with T4 PNK, and ligation with T4 ligase. Gibson assembly was performed with the Repliqa HiFi assembly mix from Quantabio, or HiFi DNA Assembly Mix (NEB).

5.6.3 Strains

E. coli strains CAG18515, JW0232, JW0233, and JW0377 were obtained from the Coli Genetic Stock Center (CGSC) at Yale University.

Proline auxotrophs generated in this work include strains SLB2001 and SLB2160. Lambda red recombination⁴⁹ was used to knock out *proA* and *proB* in DH10B (SLB2001) or BL21(DE3) (SLB2160). Briefly, pKD4 was PCR-amplified with primers delPro_fwd & delPro_rev and used to transform parent strains harboring plasmid pKD46 and expressing the recombination machinery (induced with 10 mM L-arabinose). We were unable to transform BL21 with plasmid pKD46 ourselves, so obtained strain BL21(DE3)/pKD46 from the CGSC. Transformants were plated on selective (kanamycin) LB-agar plates, and gene knockout was verified by colony PCR. pKD46 was cured growing cells at 37°C; strain SLB2160 unexpectedly retained ampicillin resistance. Proline auxotrophy of each strain was verified by comparing growth in the presence and absence of proline.

Strain DH10B was used for standard cloning operations. Electrocompetent DH5 α or 10-beta stocks were obtained from NEB, and were used for library generation.

5.6.4 DNA oligos

DNA oligos were purchased from Integrated DNA Technologies (IDT).

delPro_fwd: TAAAACGTTTCGTTTGATATCATTTTTCTAAAATTGAGTGTAGGCT
GGAGCTGCTTC

delPro_rev: GTCAATGGCCTTGTGAATCAAATGGCTACTTTTGCATATGGGAATT
AGCCATGGTCC

proS_NheI_fwd: CTAGCTAGCTCAGCCTTTAATCTGTTTCACC

proS_SacII_rev: CGCGTGTGGTAGCCGCGGCGATTGAGC

proS_SacII_fwd: GCTCAATCGCCGCGGCTACCACACGCG

proS_KpnI_rev: CCAGCTGGGTACCAAGTACTCCG

proS_PstI_fwd: AAAGTGCAGGGATTCCTGAGAAGTATGGAAAG

proS_NotI_rev: GACGAAGTGCGGCCGCGTTTCG

proS_NotI_fwd: CGAAACGCGGCCGCACTTCGTC

proS_SacI_rev: CCTGATTCGTAACGAGCTCAGCTCTTAC

del_11E11_fwd: TCGAGGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAAC
TCCG

del_11E11_rev: CTAGCGGAGTTCTGAGGTCATTACTGGATCTATCAACAGGAGT
CCC

11E11_KpnI_fwd: CGGGGTACCCTATGAGAGGATCGCATCACCATCACC

11E11_PstI_rev: AACTGCAGGCTTAATGCATAGAGCCTGAACCTGTG

11E11_RBS_fwd: CCCATTAAAGAGGAGAAATTAATGAGAGGATCGCATCA
CCATCACC

11E11_RBS_rev: TTAATTTCTCCTCTTTAATGGGGGTACCGAGCTCGAATTCGC
TAGC

AL01004_fwd: ATACCGTAGCCACCCATCGTCAGG

AL01004_rev: CGGGGTAACGCGTGTGGT

AL01005_fwd: AGCATCTTTCTGCAGGAATTCGC

AL01005_rev2: TACTCTTCCATACTTCTCAGGAATCC

SLB2110_fwd: GATTACAGGATCCTCTATGAAAATCGAAGAAGGTAAACTGGT
AATCTGG

SLB2110_rev: GTATGGATCCCGAGCCAGAGCCAGAGCCAGTCTGCGCGTCTTT
CAGGGC

SLB2111_fwd: GATTACACGGATCCGCTAGCAAAGGTGAAGAGCTGTTTGAC

SLB2111_rev: TACGGGGTACCACTAGTTTAGTCTTCGTTCCGGATCTTTGCTCAG
G

SLB2115_ins_fwd: CACCATCACCATCACGGATCCGCTAGCAAAGGTGAACGTC
TG

SLB2115-b_ins_rev: CCAAACAGCCAAGCTTGGTACCTTATTTTTCTTTCCGGAT
CTTTGCTCAGTTTAGAACG

SLB2121_rev: CCAAACAGCCAAGCTTGGTACCAC

SLB2121-b_rev: GCCAAACAGCCAAGCTTGGTAC

SLB2122_fwd: GATTACAGCATGCATGTCCAAAGGAGAAGAAGTGTACCAGG

SLB2128_fwd: GATTACAGCATGCATGAAAATCGAAGAAGGTAAACTGGTAAT
CTGG

SLB2139_ins_fwd: CATCACCATCACCATCACGGATCCGGCTCCCCGGCTGGC

SLB2139_ins_rev: GCCCGGCCGACTAGTACCCGGCGCGCTACCCTCAG

SLB2139_vec_fwd: CTGAGGGTAGCGCGCCGGTACTAGTCGGCCGGGC

SLB2139_vec_rev: GCCAGCCGGGGAGCCGGATCCGTGATGGTGTGGTGTATG

SLB2149_ins_fwd: GGTGAGAATCCAAGCTAGCTTGGCGGACTCCTGTTGATAG
ATCCAGTAATG

SLB2149_ins_rev: GCTTCCTTAGCTCCTGAAAATCTCGGATTCACGCCCTTCTCT
TTTG

SLB2149_vec_fwd: CAAAAGAGAAGGGCGTGAATCCGAGATTTTCAGGAGCTA
AGGAAGC

SLB2149_vec_rev: CATTACTGGATCTATCAACAGGAGTCCGCCAAGCTAGCTTG
GATTCTCACC

SLB2177_fwd: GGTACTAGTCGGCCGGGCTC

SLB2177_rev: GAGCCCGGCCGACTAGTACCTGGCGCGGAACCCTCG

SLB2192_fwd: ATGAGAGGATCGCATCACCATCACCATCACGGATCCAAACAA
AAAAACGGTATCAAAGCG

SLB2192_rev: TAATTTATGGCCTAAGATGTTGCC

SLB2192_LOO7_vec_fwd2: GGCAACATCTTAGGCCATAAATTATAAGGTACCAA
GCTTAATTAGCTG

SLB2192_LOO7_vec_rev: GGTGATGCGATCCTCTC

SLB2193_fwd: ATGAGAGGATCGCATCACCATCACCATCACGGATCCGAGGAT
GGATCAGTGCAAC

SLB2193_rev: TTTTGTATTATCGGCTGTGATG

SLB2193_vec_fwd: TCACAGCCGATAAACAAAAATAAGGTACCAAGCTTAATTA
GC

SLB2214_ins_fwd: CACACAGAATTCATTAAGAGGAGAAATTA ACTATGGGGG
ATATCCAGGTTCAAGTG

SLB2214_ins_rev: AAATGTTTTACAACACTCACGACAACACCAACGATGGGGCT
CGAGAGAACCGCCCTC

SLB2214_vec_fwd: ATCGTTGGTGTGTGTCGTGAGTGTGTGTA AACATTTTAAGCT
TAATTAGCTGAGCTTGGAC

SLB2214_vec_rev: GTTCACTTGAACCTGGATATCCCCATAGTTAATTTCTCCTC
TTTAATGAATTCTGTGTG

SLB2215_ins_rev: GGGCTCGAGAGAACCGCCCTC

SLB2215_vec_fwd: GTGGAAGGACAGTTAGAGGGCGGTTCTCTCGAGCCDDNT
GTTGTDDNDDNTGTTGTDDNTAAGCTTAATTAGCTGAGCTTGGAC

SLB2227_fwd: TATTGTTGTGGGGTTTGTGTATCTAAGC

SLB2228_3A_rev: AGAACCTCCGGATCCACCTGGACTACCACCTGGGGATCCAC
CTGGCTCGAGAGAACCGCC

SLB2243_ins_fwd2: CAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGCGT
TTAAGGGCACCAATAACTGC

SLB2243_ins_rev: GGTGGCACTTTTCGGGGAAATGTGCCAACTTTTGGCGAAAA
TGAGACGTTG

SLB2243_vec_fwd: CAACGTCTCATTTTCGCCAAAAGTTGGCACATTTCCCCGAA
AAGTGC

SLB2243_vec_rev2: GCAGTTATTGGTGCCCTTAAACGCTGTCAGACCAAGTTTA
CTCATATATACTTTAGATTG

SLB2322-GA_D1_fwd: GTACGCATGCTGTTACGGTTCAGTTG

SLB2327_vec_fwd: GGAGGTTCTCCAGGTGGATCCCCAGG

SLB2327_vec_rev: TGATGGTGATGGTGATGCGATCCTCTCATAGTTAATTTCTC
CTCTTTAATGAATTCTGTG

SLB2335_GA-fwd: CAACGATCAAGGCGAGTTACATGATCCC

SLB2335_GA-rev: GCACAACATGGGGGATCATGTA ACTCG

SLB2363_A-b_rev: TAAGCTTAATTAGCTGAGCTTGGAC

SLB2363_B-b_rev: AGAACCTCCGGATCCAC

SLB2371_ins_fwd: GTTGGAAGGTAGAACAGGCTCCCCAAAAAACGGGTATGG

SLB2371_ins_rev: GCGTATAATATTTGCCCATGGGGTGATGCCGGCC
 SLB2371_vec_fwd: CATCGTGGCCGGCATCACCCCATGGGCAAATATTATACGC
 SLB2371_vec_rev: CCATACCCGTTTTTTTTGGGGAGCCTGTTCTACCTTCC
 SLB2406_fwd: GCTTAATTAGCTGAGCTTGGACTCCTG
 SLB2406_rev: GCCTTGAGCCTGTTCTACCTTCC
 SLB2429_analysis_fwd: GTATCACGAGGCCCTTTTCG
 SLB2429_analysis_rev: GCTTAGATAACAACAAACCCCAACAAC
 SLB2436_4SSM-1_fwd: GAGAGGCGCTGCCGCCMNNAGAACCGGTGTCGGCTT
 G
 SLB2436_4SSM-1_rev: CATGCGTTCCCGCGAATTCCTGNNKAAANNKGCTNNK
 TCTTCCATACTTCTCAGGAATCCC
 SLB2436_4SSM-2_fwd: CACGCGTTACCCCGATACCGTAMNNACCCATCGTCAG
 GATTTGGTTACGG
 SLB2436_4SSM-2_rev: GGCGGCAGCGCCTCTC
 SLB2436_GA_vec_rev: CGGTATCGGGGTAACGCGTG
 SLB2436_GA_vec_fwd: CAGGAATTCGCGGGAACGCATG
 proS_SSM-lib1_fwd: GCCAGCACCTGGAATTCGTGAGAGGCGCTGCCGCCMNN
 GAACCGGTGTCGGCTTG
 proS_SSM-lib1_rev: AAGGAAAAAAGCGGCCGCGTTTCGGCGTCATGCGTTCCC
 GCGAATTCCTGNNKAAANNKGCTNNKTCTTCCATACTTCTCAGGAATCCC
 proS_SSM-lib2_fwd: GATTATCCCCGCGGCTACCACACGCGTTACCCCGATACC
 GTAMNNACCCATCGTCAGGATTTGGTTACGG
 proS_SSM-lib2_rev: CTCTCACGAATTCCAGGTGCTGGCGCAGAGCGGTGAAGA
 CGATGTGG
 proS_ePCR_fwd3: GGTGAGAATCCAAGCTAGCTCAGCC
 proS_ePCR_rev3: CAACTGGAACCGTAACAGCATGCGTAC
 proS_GA_vec_rev: GGCTGAGCTAGCTTGGATTCTCACC

5.6.5 g-Block gene fragments

g-Block gene fragments were purchased from Integrated DNA Technologies (IDT).

LOO#:

ATGGCATCAAAGGGAGAAGAGTTGTTTACAGGTGTAGTGCCCATCTTGGTTG
 AATTGGACGGCGATGTCAACGGACACAAATTCTCTGTACGCGGAGAGGGGGA

GGGGGACGCTACCATCGGTAACTTACGCTTAAATTTATTTGTACAACGGGA
 AAATTGCCCGTCCCCTGGCCACATTGGTAACAACCTTGACATATGGTGTGCA
 GTGTTTTAGCCGTTACCCAGATCATATGAAACGTCACGATTTCTTTAAGAGCG
 CGATGCCCCGAGGGTTACGTTTCAGGAGCGTACAATCTCGTTTAAGGATGACGG
 GAAGTATAAGACCCGCGCAGTCGTAAAGTTCGAGGGGGACACACTTGTCAAT
 CGCATTGAATTGAAGGGAAGTACTTTAAAGAGGATGGCAACATCTTAGGCC
 ATAAATTAGAGTACAACCTTTAACTCACATAACGTCTACATCACAGCCGATAA
 ACAAAAAACGGTATCAAAGCGAACTTTACAGTACGCCACAACGTTGAGGAT
 GGATCAGTGCAACTGGCGGATCATTATCAGCAGAACACGCCAATCGGGGACG
 GGCCGGTCCTTTTGCCCCACAACCACTATTTAAGCACTCAAACGGTATTGTCA
 AAAGATCCTAATGAAAAACGCGATCACATGGTACTGCTGGAATTTGTCACTG
 CCGCCGGTATTACGCATGGTATGGACGAGCTGTACAAAGGGGGTACTGGTGG
 TAGT

E6-GFP7:

ATATGCGGGGTACCCCCATTAAGAGGAGAAATTAAGTATGCGCGGCTCTCA
 CCACCACCATCATCATGGCTCTGTGCCGGGTGCGGGCGTTCCGGGAGCTGGT
 GTCCCAGGCGAAGGAGTTCCGGGTGCGGGCGTGCCCGGAGCTGGTTCTGGTA
 GTGGATCAGGGTCTGAATATAACTTCAACAGTCATAATGTGTATATCACTGCA
 GACGGTTCAGGATCGGGGTCCGAATACAATTTCAATTCCATAATGTCTACAT
 CACAGCTGACTGAGCCTGCAGGCATGCAAGCTTGATTACA

E6-GFP8:

ATATGCGGGGTACCCCCATTAAGAGGAGAAATTAAGTATGCGTGGTAGTCA
 CCACCATCATCATCATGGTTCAGTTCAGGTGCGGGTGTACCCGGCGCAGGT
 GTTCCAGGCGAAGGTGTGCCGGGGGCGGGGGTCCCGGCGCGGGCAGTGGTT
 CCGTTCGGTTCAAATGGGATTAAGCTAATTTTACAGTACGCCACAATGTC
 GGGAGTGGTAGCGGCTCGAACGGCATTAAAGCAAATTTTACAGTCCGTCATA
 ATGTCTAAGCCTGCAGGCATGCAAGCTTGATTACA

DHR14:

ATGAGAGGATCGCATCACCATCACCATCACATGGATAGCGAGGAAGTGAATG
 AACGTGTAAAACAACCTGGCAGAGAAAGCAAAGGAAGCAACCGATAAAGAAG
 AGGTGATTGAAATTGTAAGAGAGTTAGCCGAATTGGCCAAGCAGAGTACTGA
 CTCCGAATTAGTCAACGAGATTGTGAAGCAATTAGCAGAAGTGGCGAAGGAG
 GCGACAGATAAGGAGTTGGTAATTTACATCGTGAAGATCTTGGCTGAATTAG
 CGAAGCAAAGCACCGACAGCGAACTGGTAAACGAGATTGTAAAACAGCTTG
 CGGAGGTGGCTAAAGAGGCAACGGACAAAGAGTTAGTGATTTATATCGTAAA
 GATCCTGGCTGAGTTAGCCAAGCAATCCACTGATTCAGAGTTGGTTAACGAA
 ATTGTAAGCAGCTTGAAGAGGTCGCCAAAGAGGCTACTGACAAAGAAGT
 GTAGAACACATTGAAAAGATCCTGGAGGAGCTTAAAAGCAGTCCACAGAC
 GTTTGGCTGGAGGGAGGTTCTCCAGGTGGATCCCCA

5.6.6 Plasmids and cloning

pBAD33_mWasabi: the mWasabi gene was amplified from plasmid pKPY515 (Ref. 50) and installed in the pBAD33 backbone using KpnI and HindIII cut sites.

pQE80PI-proS was previously described.¹⁹

pQE80_11-E-11_proS: To facilitate site-saturation mutagenesis library creation, cut sites NotI and SacII were installed into the *proS* gene of pQE80β11(Eβ11)-proS.¹⁹ These cut sites flank the library residues of interest, but introduce silent mutations into the *proS* coding sequence. We used an overlap extension PCR approach: PCR fragment 1 (amplified with primers proS_NheI_fwd & proS_SacII_rev) was stitched together with PCR fragment 2 (proS_SacII_fwd & proS_KpnI_rev), then ligated into the vector backbone using a restriction enzyme cloning approach (NheI and KpnI cut sites) to install the SacII cut site. PCR fragment 3 (proS_PstI_fwd & proS_NotI_rev) was stitched together with PCR fragment 4 (proS_NotI_fwd & proS_SacI_rev) and ligated into the digested vector backbone (PstI & SacI cut sites) to install the NotI cut site.

ptac_GFP(1-10) was previously described.¹⁹

pQE80_proS: The 11-E-11 gene was excised from the plasmid pQE80_11-E-11_proS using cut sites XhoI and EcoRI; this excised sequence was replaced with a short DNA fragment resulting from annealing the oligos del_11E11_fwd and del_11E11_rev.

pBAD33_11-E-11: The 11-E-11 gene was amplified from plasmid pQE80_11-E-11_proS using primers 11E11_KpnI_fwd & 11E11_PstI_rev and installed into the pBAD33 backbone using restriction enzyme cloning (KpnI & PstI cut sites). The RBS was then installed with a one-piece Gibson Assembly approach with the DNA fragment resulting

from PCR amplification of the vector backbone using primers 11E11_RBS_fwd & 11E11_RBS_rev.

pQE80_GFP1-10: GFP1-10 was amplified from ptac_GFP(1-10) with primers SLB2122_fwd & SLB2121_rev and installed into the pQE-80L backbone using a restriction enzyme approach (HindIII and SphI cut sites).

pQE80_MBP-GFP1-10: The MBP gene was first installed at the N-terminus of GFP1-10 in vector ptac_GFP(1-10): MBP was amplified from plasmid pET His6 MBP TEV LIC (a gift from Scott Gradia, Addgene plasmid #29656) with primers SLB2110_fwd & SLB2110_rev, then installed at the BamHI cut site in ptac_GFP(1-10). The MBP-GFP1-10 fusion protein gene was transferred to the pQE-80L backbone using the HindIII and SphI cut sites after amplification with SLB2128_fwd & SLB2121-b_rev. A similar approach was used to install NusA, TrxA, and SUMO at the N-terminus of GFP1-10.

pQE80_MBP-GFP1-10_proS: We used a Gibson Assembly approach to install the *proS* gene into the pQE80_MBP-GFP1-10 backbone, with primers SLB2149_ins_fwd & SLB2149_ins_rev (*proS* amplification), and SLB2149_vec_fwd & SLB2149_vec_rev (backbone amplification).

pBAD33_X144-GFP11 was cloned using a Gibson Assembly approach. The XTEN144 gene was amplified from pX_X144 (a gift from Peter B. Rapp) with primers SLB2139_ins_fwd & SLB2139_ins_rev, and replaced the N-terminal GFP11-elastic fragment in the pBAD33_11-E-11 vector (primers SLB2139_vec_fwd & SLB2139_vec_rev).

pBAD33_X72-GFP11: The C-terminal half of the XTEN144 gene in pBAD33_X144-GFP11 was removed with a one-piece Gibson Assembly approach, after amplification with primers SLB2177_fwd & SLB2177_rev.

ptac_sn1-10: β -strands 1-10 from super-negative GFP³² were amplified from pET-6xHis(-30)GFP (a gift from David Liu, Addgene plasmid #62936) with primers SLB2111_fwd & SLB2111_rev. The amplicon was installed in place of GFP1-10 in vector ptac_GFP(1-10) using restriction enzyme cloning (cut sites BamHI & KpnI).

ptac_sp1-10: β -strands 1-10 from super-positive GFP³² were amplified from pET-6xHis(pos36)GFP (a gift from David Liu, Addgene plasmid #62937) with primers SLB2115_ins_fwd & SLB2115-b_ins_rev. The amplicon was installed in place of GFP1-10 in vector ptac_GFP(1-10) using restriction enzyme cloning (cut sites BamHI & KpnI).

pQE80_LOO7_proS: A g-Block gene fragment (LOO#) corresponding to the sequence of the circularly permuted GFP was circularized by ligation. It was amplified by primers SLB2192_fwd & SLB2192_rev to give the large LOO7 GFP fragment. This GFP fragment replaced the MBP-GFP1-10 gene in vector pQE80_MBP-GFP1-10 by Gibson Assembly: the backbone fragment was amplified with primers SLB2192_LOO7_vec_fwd2 & SLB2192_LOO7_vec_rev.

pQE80_LOO8_proS: A g-Block gene fragment (LOO#) corresponding to the sequence of the circularly permuted GFP was circularized by ligation. It was amplified by primers SLB2193_fwd & SLB2193_rev to give the large LOO8 GFP fragment. This GFP fragment replaced the MBP-GFP1-10 gene in vector pQE80_MBP-GFP1-10 by Gibson Assembly:

the backbone fragment was amplified with primers SLB2193_vec_fwd & SLB2192_LOO7_vec_rev.

pBAD33_E-7: A g-Block gene fragment (E6-GFP7) was designed to contain two copies of the GFP7 strand downstream of a six-mer elastin fragment. It was installed in the pBAD33 vector using restriction enzyme cloning (KpnI and HindIII cut sites).

pBAD33_E-8: A g-Block gene fragment (E6-GFP8) was designed to contain two copies of the GFP8 strand downstream of a six-mer elastin fragment. It was installed in the pBAD33 vector using restriction enzyme cloning (KpnI and HindIII cut sites).

pQE80_Top7-P-TC_proS: The initial design of the Top7-TC gene included one proline codon between the Top7 gene and the TC motif HRWCCRECKTF.⁵¹ The Top7 gene was amplified with primers SLB2214_ins_fwd & SLB2214_ins_rev from plasmid Top7 K39E K40E V48V K55E, a gift from David Baker (Addgene plasmid # 12464). It replaced the 11-E-11 gene in pQE80_11-E-11_proS by Gibson Assembly (SLB2214_vec_fwd & SLB2214_vec_rev).

pQE80_Top7-TC3_proS: We installed a longer proline-containing linker between Top7 and the TC motif TC3 to increase screening stringency. The sorted TC library variant pQE80_Top7-P-TC3_proS was amplified with primers SLB2227_fwd & SLB2228_3A_rev. The linear PCR product was phosphorylated, then ligated via blunt end ligation.

pQE80-c_Top7-TC3_proS: A Gibson Assembly approach was used to replace the ampR cassette with that of CAT from pBAD33. pQE80_Top7-TC3_proS was amplified with SLB2243_vec_fwd & SLB2243_vec_rev2; the CAT gene from pBAD33 was amplified

with SLB2243_ins_fwd2 & SLB2243_ins_rev. This plasmid was used for Top7-TC3 expression in strain SLB2160, which retained ampicillin resistance after lambda red recombination.

pQE80_DHR14-TC3_proS: Gibson Assembly was used to replace the Top7 gene with DHR14. pQE80_Top7-TC3_proS was amplified with primers SLB2327_vec_fwd & SLB2327_vec_rev, and was assembled with the codon-optimized DHR14 g-Block. A similar approach was used to generate plasmids corresponding to the other proline-free genes screened for toxicity and FIAsh labeling.

pQE80_DHR14-TC3: The DHR14-TC3 gene (from plasmid pQE80_DHR14TC3_proS) was installed into the pQE-80L backbone by restriction enzyme cloning (XhoI & HindIII cut sites).

pQE80_DHR14-3P_proS: The TC3 motif was removed from the DHR14 gene by a blunt-end ligation approach after PCR amplification of pQE80_DHR14-TC3_proS with primers SLB2363_A-b_rev & SLB2363_B-b_rev.

pQE80_DHR14-TC3_proS-ara: We cloned an arabinose-inducible version of the *proS* gene through a three-part Gibson Assembly approach. The AraC fragment was amplified from pBAD33 with primers SLB2371_ins_fwd & SLB2371_ins_rev. pQE80_DHR14-TC3_proS was amplified as two fragments using the following primer sets: SLB2371_vec_fwd / AmpR-GA_rev and AmpR-GA_fwd / SLB2371_vec_rev.

M157Q & C443G point mutations: A blunt-end ligation approach was used to install the desired point mutations into the *proS* gene. Template plasmids were amplified with

AL01005_fwd & AL01005_rev2 (M157Q), or AL01004_fwd & AL01004_rev (C443G).

The linear PCR product was phosphorylated, then ligated via blunt end ligation.

5.6.7 Library generation

TC motif library: The 8-mer randomized amino acid sequence XCCXXCCX was designed, in which DDN (D = G/A/T; N = A/T/G/C) codons for each X residue were used. We installed this TC motif library at the C-terminus of the Top7 gene using a Gibson Assembly approach: the Top7 gene was amplified with primers SLB2214_ins_fwd & SLB2215_ins_rev from plasmid Top7 K39E K40E V48V K55E, (a gift from David Baker, Addgene plasmid # 12464) and pQE80_11-E-11_proS was amplified with SLB2215_vec_fwd & SLB2214_vec_rev.

5-site saturation *proS* mutagenesis: The *proS* gene was amplified from plasmid pQE80_11-E-11_proS in two fragments. Insert 1 contains the NNK-randomized positions 157, 159, 161, and 202 (K = G/T; N = A/T/G/C), and was amplified with primers proS_SSM-lib1_fwd & proS_SSM-lib1_rev. Insert 2 contains randomization at position 443, and was amplified with primers proS_SSM-lib2_fwd & proS_SSM-lib2_rev. Insert 1 and insert 2 were stitched together using overlap extension PCR. The resulting PCR product was installed into vector pQE80_Top7-TC3_proS using a restriction enzyme approach (SacII & NotI cut sites). We obtained approximately 2×10^7 transformants, and NNK randomization was verified by Sanger sequencing.

4-site saturation *proS* mutagenesis: We constructed our 4SSM library through two approaches. In the original restriction enzyme approach, the *proS* gene was amplified from plasmid pQE80_DHR14-TC3_proS with primers proS_SSM-lib1_fwd & proS_SSM-

lib2_rev, leading to NNK-randomization at positions 157, 159, 161, and 443. The resulting PCR product was installed in the pQE80_DHR14-TC3_proS backbone using restriction enzyme cloning (SacII and NotI cut sites). We obtained approximately 4×10^5 transformants.

In the second approach, we performed a four-part Gibson Assembly with fragments resulting from the PCR amplifications of pQE80_DHR14-TC3_proS using the following primer sets: (1) SLB2436_4SSM-1_fwd & SLB2436_4SSM-1_rev; (2) SLB2436_4SSM-2_fwd & SLB2436_4SSM-2_rev; (3) SLB2335_GA-fwd & SLB2436_GA_vec_rev; and (4) SLB2436_GA_vec_fwd & SLB2335_GA-rev. In an initial test used to measure plasmid stability, we obtained approximately 2×10^5 transformants.

Error-prone PCR: We constructed our ePCR library through two approaches. In the original restriction enzyme approach, the *proS* gene from plasmid pQE80_DHR14-TC3_proS was amplified with primers proS_ePCR_fwd3 & proS_ePCR_rev3 with the Taq DNA polymerase (NEB) and in the presence of 0-400 μM MnCl_2 . The PCR product was installed into pQE80_DHR14-TC3_proS using restriction enzyme cloning (NheI & SphI cut sites) and an extended 16 hour ligation reaction at 16°C with T4 ligase. Increasing MnCl_2 concentrations yielded increasing *proS* mutation rates, as expected (Table 5.S1). We chose to sort the library generated under 200 μM MnCl_2 conditions (2.13 ± 1.87 mutations per *proS* gene), and obtained 1×10^6 transformants.

In the second approach, we performed a three-part Gibson Assembly. Fragments corresponding to the plasmid backbone were amplified with Q5 DNA polymerase and the following primer sets: (1) SLB2335_GA-fwd & ProS_GA_vec_rev; and (2) SLB2322-

GA_D1_fwd & SLB2335_GA-rev. The *proS* gene was amplified with Taq polymerase in the presence of 200 μ M MnCl₂ with primers proS_ePCR_fwd3 & proS_ePCR_rev3. In an initial test used to measure plasmid stability, we obtained approximately 3×10^4 transformants.

5.6.8 Nucleotide and amino acid sequences

mWasabi:

ATGAGAGGATCGCATCACCATCACCATCACGGATCCATGGTGAGCAAGGGCG
 AGGAGACCACAATGGGCGTAATCAAGCCCGACATGAAGATCAAGCTGAAGA
 TGGAGGGCAACGTGAATGGCCACGCCTTCGTGATCGAGGGCGAGGGCGAGG
 GCAAGCCCTACGACGGCACCAACACCATCAACCTGGAGGTGAAGGAGGGAG
 CCCCCCTGCCCTTCTCCTACGACATTCTGACCACCGCGTTCAGTTACGGCAAC
 AGGGCCTTCACCAAGTACCCCGACGACATCCCCAACTACTTCAAGCAGTCCT
 TCCCCGAGGGCTACTCTTGGGAGCGCACCATGACCTTCGAGGACAAGGGCAT
 CGTGAAGGTGAAGTCCGACATCTCCATGGAGGAGGACTCCTTCATCTACGAG
 ATACACCTCAAGGGCGGAGAACTTCCCCCCAACGGCCCCGTGATGCAGAAGG
 AGACCACCGGCTGGGACGCCTCCACCGAGAGGATGTACGTGCGCGACGGCGT
 GCTGAAGGGCGACGTCAAGATGAAGCTGCTGCTGGAGGGGCGGCGGCCACCA
 CCGCGTTGACTTCAAGACCATCTACAGGGCCAAGAAGGCGGTGAAGCTGCCC
 GACTATCACTTTGTGGACCACCGCATCGAGATCCTGAACCACGACAAGGACT
 ACAACAAGGTGACCGTTTACGAGATCGCCGTGGCCCGCAACTCCACCGACGG
 CATGGACGAGCTGTACAAGTAA

MRGSHHHHHHGSMSVSKGEETTMGVIKPDMKIKLKMENGVNGHAFVIEGEGEG
 KPYDGTNTINLEVKEGAPLPFSYDILTTAFSYGNRAFTKYPDDIPNYFKQSFPEGY
 SWERTMTFEDKGIVKVKSDISMEEDSFIYEIHLKGENFPPNGPVMQKETTGWAS
 TERMYVRDGVKGDVKMKLLLEGGGHRVDFKTIYRAKKAVKLPDYHFVDHR
 IEILNHDKDYNKVTVYEIAVARNSTDGMDELYK

GFP11-Elastin-GFP11:

ATGAGAGGATCGCATCACCATCACCATCACGGATCCCGGCCGGGCTCTGGTT
 CTGGTTCAGGCTCTCGTGACCATATGGTTCTTCATGAGTACGTAAATGCTGCT
 GGCATCACAGGGTCGACTGTGCCGGGTGCGGGCGTTCCGGGAGCTGGTGTCC
 CAGGCGAAGGAGTTCGGGTGCGGGCGTGCCCGGAGCTGGTACTAGTCGGCC
 GGGCTCTGGTTCTGGTTCAGGCTCTCGTGACCATATGGTTCTTCATGAGTACG
 TAAATGCTGCTGGCATCACAGGTTTCAGGCTCTATGCATTAA

MRGSHHHHHHGSRPGS SSGSGSRDHMVLHEYVNAAGITGSTVPGAGVPGAGVP
 GEGVPGAGVPGAGTSRPGS SSGSGSRDHMVLHEYVNAAGITGSGSMH

GFP1-10:

ATGAGAGGATCGCATCACCATCACCATCACGGATCCATGTCCAAAGGAGAAG
 AACTGTTTACCGGCGTTGTGCCAATTTTGGTTGAACTCGATGGTGTATGTCAAC
 GGACATAAGTTCTCAGTGAGAGGCGAAGGAGAAGGTGACGCCACCATTGGA
 AAATTGACTCTTAAATTCATCTGTACTACTGGTAAACTTCCTGTACCATGGCC
 GACTCTCGTAACAACGCTTACGTACGGAGTTCAGTGCTTTTCGAGATACCCAG
 ACCATATGAAAAGACATGACTTTTTTAAGTCGGCTATGCCTGAAGGTTACGTG
 CAAGAAAGAACAATTCGTTCAAAGATGATGGAAAATATAAAACTAGAGCA
 GTTGTTAAATTTGAAGGAGATACTTTGGTTAACCGCATTGAACTGAAAGGAA
 CAGATTTTAAAGAAGATGGTAATATTCTTGGACACAAACTCGAATACAATTTT
 AATAGTCATAACGTATACATCACTGCTGATAAGCAAAGAACGGAATTAAG
 CGAATTTACAGTACGCCATAATGTAGAAGATGGCAGTGTTCAACTTGCCGA
 CCATTACCAACAAAACACCCCTATTGGAGACGGTCCGGTACTTCTTCCTGATA
 ATCACTACCTCTCAACACAAACAGTCCTGAGCAAAGATCCAAATGAAAAATA
 A

MRGSHHHHHHGSMSKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATIGK
 LTLKFICTTGKLPVPWPTLVTTLTYGVCFSRYPDHMKRHDFFKSAMPEGYVQE
 RTISFKDDGKYKTRAVVKFEGDTLVNRIELKGTDFKEDGNILGHKLEYNFNSHN
 VYITADKQKNGIKANFTVRHNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLST
 QTVLSKDPNEK

XTEN144-GFP11:

ATGAGAGGATCGCATCACCATCACCATCACGGATCCGGCTCCCCGGCTGGCA
 GCCCGACCAGCACTGAAGAGGGCACGAGCGAGTCGGCGACCCCGGAGTCTG
 GTCCGGGCACCTCCACCGAACCGTCTGAGGGCAGCGCACCGGGTAGCCCGGC
 CGGTAGCCCTACCAGCACCGAAGAGGGTACCAGCACGGAACCGAGCGAAGG
 CTCGGCACCGGGTACGAGCACCGAGCCGTCCGAGGGTTCGCGCCAGGTACC
 AGCGAGAGCGCAACGCCGGAGTCCGGTCCGGGCAGCGAACCAGCGACCAGC
 GGCAGCGAAACGCCGGGTTACAGAGCCGGCGACGAGCGGTAGCGAGACTCCG
 GGCAGCCCGGCTGGTAGCCCGACGTCCACCGAAGAAGGCACCAGCGAAAGC
 GCCACCCCGGAGAGCGGTCCTGGTACGTCTACCGAGCCATCTGAGGGTAGCG
 CGCCGGGTACTAGTCGGCCGGGCTCTGGTTCTGGTTCAGGCTCTCGTGACCAT
 ATGGTTCTTCATGAGTACGTAAATGCTGCTGGCATCACAGGTTTCAGGCTCTAT
 GCATTAA

MRGSHHHHHHGS SPAGSPTSTEEGTSESATPESGPGTSTEPSEGSAPGSPAGSPT
 STEEGTSTEPSEGSAPGTSTEPSEGSAPGTSESATPESGPGSEPATSGSETPGSEPAT
 SGSETPGSPAGSPTSTEEGTSESATPESGPGTSTEPSEGSAPGTSRPGSGSGSGSRD
 HMLVHEYVNAAGITGSGSMH

Prolyl-tRNA synthetase (ProRS):

ATGCGTACTAGCCAATACCTGCTCTCCACTCTCAAGGAGACACCTGCCGACG
 CCGAGGTGATCAGCCATCAGCTGATGCTGCGCGCCGGGATGATCCGCAAGCT
 GGCCTCCGGGTTATATACCTGGCTGCCGACCGGCGTGCGCGTTCTGAAAAAA
 GTCGAAAACATCGTGCGTGAAGAGATGAACAACGCCGGTGCGATCGAGGTGT
 CGATGCCGGTGGTTCAGCCAGCCGATTTGTGGCAAGAGAGTGGTCGTTGGGA
 ACAGTACGGTCCGGAACCTGCTGCGTTTTGTTGACCGTGGCGAGCGTCCGTTCCG
 TACTCGGCCCAACTCATGAAGAAGTTATCACTGACCTGATTCGTAACGAGCTC
 AGCTCTTACAAACAGCTGCCGCTGAACTTCTATCAGATCCAGACCAAGTTCC
 GCGACGAAGTGCGGCCGCGTTTTCGGCGTCATGCGTTCCCGCGAATTCCTGAT
 GAAAGATGCTTACTCTTTCCATACTTCTCAGGAATCCCTGCAGGAAACCTACG
 ATGCAATGTATGCGGCCTACAGCAAAATCTTCAGCCGCATGGGGCTGGATTT
 CCGCGCCGTACAAGCCGACACCGGTTCTATCGGCGGCAGCGCCTCTCACGAA

TTCCAGGTGCTGGCGCAGAGCGGTGAAGACGATGTGGTCTTCTCCGACACCT
CTGACTATGCAGCGAACATTGAACTGGCAGAAGCTATCGCGCCGAAAGAACC
GCGCGCTGCTGCTACCCAGGAAATGACGCTGGTTGATACGCCGAACGCGAAA
ACCATCGCGGAACTGGTTGAACAGTTCAATCTGCCGATTGAGAAAACGGTTA
AGACTCTGCTGGTTAAAGCGGTTGAAGGCAGCAGCTTCCCGCAGGTTGCGCT
GCTGGTGCAGCGGTGATCACGAGCTGAACGAAGTTAAAGCAGAAAACTGCC
GCAGGTTGCAAGCCCGCTGACTTTCGCGACCGAAGAAGAAATTCGTGCCGTG
GTTAAAGCCGGTCCGGGTTCACTGGGTCCGGTAAACATGCCGATTCCGGTGG
TGATTGACCGTACCGTTGCGGCGATGAGTGATTTGCTGCTGGTGCTAACATC
GATGGTAAACACTACTTCGGCATCAACTGGGATCGCGATGTCGCTACCCCGG
AAGTTGCAGATATCCGTAACGTGGTGGCTGGCGATCCAAGCCCGGATGGCCA
GGGTAGGCTGCTGATCAAACGTGGTATCGAAGTTGGTCACATCTTCCAGCTG
GGTACCAAGTACTCCGAAGCACTGAAAGCCTCCGTACAGGGTGAAGATGGCC
GTAACCAAATCCTGACGATGGGTTGCTACGGTATCGGGGTAACGCGTGTGGT
AGCCGCGGCGATTGAGCAGAACTACGACGAACGAGGCATCGTATGGCCTGA
CGCTATCGCGCCGTTCCAGGTGGCGATTCTGCCGATGAACATGCACAAATCC
TTCCGCGTACAAGAGCTTGCTGAGAACTGTACAGCGAACTGCGTGCACAAG
GTATCGAAGTGCTGCTGGATGACCGCAAAGAGCGTCCGGGCGTGATGTTTGC
TGATATGGAAGTATCGGTATTCCGCACACTATTGTGCTGGGCGACCGTAACC
TCGACAACGACGATATCGAATATAAATATCGTCGCAACGGCGAGAAACAGTT
AATTAAGACTGGTGACATCGTTCGAATATCTGGTGAACAGATTAAAGGCTGA
MRTSQYLLSTLKETPADAEVISHQLMLRAGMIRKLASGLYTWLPTGVRVLKKVE
NIVREEMNAGAIEVSMPVVQPADLWQESGRWEQYGPELLRFVDRGERPFVLG
PTHEEVITDLIRNELSSYKQLPLNFYQIQTKFRDEVPRFGVMRSREFLMKDAYSF
HTSQESLQETYDAMYAAYSKIFSRMGLDFRAVQADTGSIGGSASHEFQVLAQSG
EDDVVFSDTSDYAANIELAEAIAPKEPRAAATQEMTLVDTPNAKTIAELVEQFNL
PIEKTVKTL LVKAVEGSSFPQVALLVRGDHELNEVKAELPQVASPLTFATEEEIR
AVVKAGPGSLGPVNMPIPVVIDRTVAAMSDFAAGANIDGKH YFGINWDRDVAT
PEVADIRNVVAGDPSPDGQGRLLIKRGIEVGHIFQLGTYSEALKASVQGEDGRN
QILTMGCYGIGVTRVVAAAIEQNYDERGIVWPDAIAPFQVAILPMNMHKSFRVQ

ELAEKLYSELRAQGIEVLLDDRKERPGVMFADMELIGIPHTIVLGDRNLDNDIE
YKYRRNGEKQLIKTGDIVEYLVKQIKG

Top7-TC3:

ATGGGGGATATCCAGGTTCAAGTGAACATTGACGATAATGGTAAAACTTCG
ATTATACTTATAACCGTGACCACAGAATCGGAATTACAGAAGGTTTTGAACGA
ACTGAAAGACTACATCGAAGAACAGGGGGCTAAACGCGCGCGCATCTCGATT
ACCGCTCGTACGGAAAAAGAAGCAGAAAAATTTGCCGCGATTTTGATTAAAG
TTTTCGCCGAACTGGGTTATAATGATATTAACGTTACGTGGGACGGCGATACG
GTGACAGTGGAAGGACAGTTAGAGGGCGGTTCTCTCGAGCCCCATCGTTGGT
GTTGTCGTGAGTGTTGTAAAACATTTTAA

MGDIQVQVNIDDNGKNFDYTYTVTTESELQKVLNELKDYIEEQAKRARISITAR
TEKEAEKFAAILIKVFAELGYNDINVTWDGDTVTVEGQLEGGSLPHRWCCREC
CKTF

DHR14-TC3:

ATGAGAGGATCGCATCACCATCACCATCACATGGATAGCGAGGAAGTGAATG
AACGTGTAAAACAACCTGGCAGAGAAAGCAAAGGAAGCAACCGATAAAGAAG
AGGTGATTGAAATTGTAAAAGAGTTAGCCGAATTGGCCAAGCAGAGTACTGA
CTCCGAATTAGTCAACGAGATTGTGAAGCAATTAGCAGAAGTGGCGAAGGAG
GCGACAGATAAGGAGTTGGTAATTTACATCGTGAAGATCTTGGCTGAATTAG
CGAAGCAAAGCACCGACAGCGAACTGGTAAACGAGATTGTAAAACAGCTTG
CGGAGGTGGCTAAAGAGGCAACGGACAAAGAGTTAGTGATTTATATCGTAAA
GATCCTGGCTGAGTTAGCCAAGCAATCCACTGATTCAGAGTTGGTTAACGAA
ATTGTAAAGCAGCTTGAAGAGGTCGCCAAAGAGGCTACTGACAAAGAACTG
GTAGAACACATTGAAAAGATCCTGGAGGAGCTTAAAAGCAGTCCACAGAC
GGTTGGCTGGAGGGAGGTTCTCCAGGTGGATCCCAGGTGGTAGTCCAGGTG
GATCCGGAGGTTCTTATTGTTGTGGGGTTTGTGTATCTAA

MRGSHHHHHHMDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEKQSTDSE
LVNEIVKQLAEVAKEATDKELVIYIVKILAEKQSTDSELVNEIVKQLAEVAKE

ATDKELVIYIVKILAEELAKQSTDSELVNEIVKQLEEVAKEATDKELVEHIEKILEEL
KKQSTDGWLEGGSPGGSPGGSPGGSGGSYCCGVCCI

5.6.9 Non-canonical proline incorporation and protein expression

A general protocol for ncPro incorporation is described here; exact conditions for some experiments may differ. A proline auxotroph strain of *E. coli* (here, usually strain CAG18515) carrying a plasmid encoding a protein of interest (POI) and overexpressing the *EcProRS* (or a point mutant thereof) was grown from a single colony to stationary phase overnight in rich medium supplemented with ampicillin (or the appropriate alternative antibiotic). The culture was diluted 1:100 into 1x 20aa M9 medium of the following composition: 8.5 mM NaCl, 18.7 mM NH₄Cl, 22 mM KH₂PO₄, 47.8 mM Na₂HPO₄, 0.1 mM CaCl₂, 1 mM MgSO₄, 3 mg L⁻¹ FeSO₄, 1 μg L⁻¹ trace metals [Cu²⁺, Mn²⁺, Zn²⁺, MoO₄²⁻], 35 mg L⁻¹ thiamine HCl, 10 mg L⁻¹ biotin, 20 mM D-glucose, 100 mg L⁻¹ ampicillin, 50 mg L⁻¹ of each of the twenty canonical L-amino acids. Cultures were grown at 37°C with shaking until they reached late-exponential phase (OD₆₀₀ ~0.8), at which point they were subjected to a medium shift: cells were pelleted via centrifugation (4,000 g for 6 min, 4°C), and washed twice with ice-cold 0.9% NaCl. Washed cells were re-suspended in 1.25x 19aa M9 medium (1.25x-concentrated M9 medium, without L-proline). After cells were incubated for 30 min at 37°C to deplete L-proline, 0.3 M NaCl and 0.5 mM ncPro were added to each culture. Cells were incubated for 30 min at 37°C to allow ncPro uptake, and expression of the POI was induced with IPTG.

For staggered expression of GFP1-10/11 fragments, expression of the GFP1-10 fragment was induced with IPTG at mid-log phase (OD₆₀₀ ~0.4) at 25°C before the medium shift

into 1.25x 19aa M9 medium; at this stage, glycerol replaced glucose as the carbon source for araBAD promoter use. After proline depletion and ncPro uptake steps, expression of the GFP11 strand was induced with 0.1% arabinose.

5.6.10 Quantification of ncPro incorporation

ncPro residues were incorporated into proinsulin under the general conditions described above using plasmid pQE8PI-proS. Proinsulin expression was induced with 1 mM IPTG for 2.5 h, after which cells were harvested by centrifugation and stored at -80°C until further processing.

Cell pellets were thawed and lysed with B-PER complete (Thermo Scientific) for 1 h at room temperature. Lysates were cleared by centrifugation (10,000 g, 10 min), and the supernatant was discarded. The insoluble inclusion body fraction was washed once with wash buffer (2 M urea, 20 mM tris, 1% triton X-100, pH 8.0) and twice with ddH₂O. The inclusion body was resuspended in solubilization buffer (8 M urea, 300 mM NaCl, 50 mM NaH₂PO₄, pH 8.0) and incubated for 1 h at RT, or overnight at 4°C. Insoluble debris was removed by centrifugation (10,000 g, 10 min) and discarded. The supernatant (now containing solubilized proinsulin) was analyzed by SDS-PAGE and MALDI-TOF mass spectrometry. The presence of a distinct proinsulin band by SDS-PAGE usually corresponded to high incorporation efficiencies by mass spectrometry.

To quantify ncPro incorporation, we digested proinsulin with the peptidase Glu-C and analyzed the peptide fragment ³³RGFFYTPKTRRE by MALDI-TOF. Crude solubilized proinsulin was reduced (5 mM DTT, 55°C, 20 min) and alkylated (15 mM iodoacetamide, RT, 15 min) prior to 10-fold dilution in to 100 mM NH₄HCO₃, pH 8.0 (100 μM total

volume). 0.6 μL of 0.5 $\mu\text{g } \mu\text{L}^{-1}$ Glu-C (Promega) was added, and samples were digested for 2.5 h at 37°C. Digestion was halted by the addition of 10 μL of 5% TFA, and peptides were desalted using ZipTip with C₁₈ resin (MilliporeSigma) according to the manufacturer's protocols. Desalted peptides (in 50% acetonitrile, 0.1% TFA) were diluted 3:1 into α -CN matrix solution (α -cyanohydroxycinnamic acid in 50% ACN, 0.1% TFA) and analyzed by MALDI-TOF MS. Analog incorporation was calculated by comparing total area under the curve (AUC) of the ion corresponding to the ncPro form of the peptide, with total AUC of that of the canonical proline peptide ($m/z = 1558$).

5.6.11 Western blot analysis of protein solubility

1 mL samples were centrifuged (14 kg, 1 min) then stored at -80°C. Cells were lysed with B-PER Complete (Thermo Fisher, 5 mL/g cell pellet) for 1 h at room temperature, then centrifuged (14 kg, 10 min). The supernatant was removed as the soluble fraction. The inclusion body fraction was washed once with Triton X wash buffer and twice with ddH₂O. The washed inclusion body was resuspended in Ni-NTA binding buffer for 1 h at room temperature, then centrifuged to remove insoluble debris. The supernatant was collected as the inclusion body fraction. 9 μL of solubilized inclusion body was mixed with 3 μL SDS-PAGE loading buffer and heated to 95°C for 5-10 min, then loaded onto a Nu-PAGE bis-tris 4-12% pre-cast gel. Proteins were transferred onto a nitrocellulose membrane by the iBlot 2 Dry Blotting System (Thermo Fisher). The blot was blocked with a 5% powdered milk solution in PBST and stained with an Alexa488-conjugated anti-hexahistidine antibody (4°C, overnight). The western blot was washed three times with PBST, then imaged on a Typhoon Trio Variable Mode Imager (GE Healthcare) with a 488 nm laser and 520 nm band-pass filter.

5.6.12 *FLAsH labeling (optimized conditions)*

TC-tagged proteins were expressed as the POI under the general ncPro incorporation conditions described above, except for the following changes. 100 μ M FLAsH-EDT₂ and 200 μ M EDT were added (as 500x stocks dissolved in DMSO) after the medium shift and resuspension in 1.25x 19aa M9 medium. Samples were kept in the dark after the addition of FLAsH. After proline depletion and ncPro uptake steps, Top7-TC3 or DHR14-TC3 expression was induced with 0.5 mM IPTG for 2.5 h at 37°C, then prepared for flow cytometry (see below).

5.6.13 *Flow cytometry and FACS*

In preparation for analysis by flow cytometry, samples were centrifuged, washed twice with PBS, and filtered through 40 μ m nylon strainers (Corning). For cells labeled with FLAsH, we found that washing once with PBSDE (33 mM EDT, 3% v/v DMSO in PBS), once with PBS (3% v/v DMSO in PBS), then resuspending samples in PBS before filtering samples modestly reduced background FLAsH labeling and improved resolution between –Proline and +Proline controls. Cells were analyzed on a MoFlo XDP cell sorter (Beckman Coulter) using a 488 nm laser and 530/40 nm bandpass filter.

Events with the desired fluorescence characteristics were sorted: for libraries, these were the brightest ~1%; doublets were excluded from sorting by gating on FL1-Hight vs FL1-Area. Cells were sorted into rich medium and rescued for one hour at 37°C, then diluted into selective medium and grown overnight at 37°C. Samples were either diluted the next day into M9 medium from this overnight culture for subsequent rounds of flow cytometry analysis and FACS, or stored at -80°C in 25% glycerol stocks until further analysis. Rescue

efficiency was determined by plating cells immediately after the 1 h rescue step onto selective agar plates, and was defined as CFUs / events sorted.

Alternatively, cells were sorted into empty tubes, then centrifuged (20,000 g, 10 min). The supernatant was carefully removed, and plasmid DNA was isolated using a ZymoPURE plasmid miniprep kit. The manufacturer's protocols were generally followed, though buffer volumes were scaled by a factor of 2/7, and Zymo-Spin I columns (rather than Zymo-Spin II columns) were used. Eluted plasmid DNA was then used to transform new stocks of *E. coli*.

5.6.14 Replica plating

E. coli strain CAG18515 harboring plasmid pQE80PI-proS, pQE80PI-proS-M157Q, or pQE80PI were plated onto LB agar plates containing ampicillin and grown until large colonies formed. This master plate was imaged (EPSON scanner), then colonies were transferred to sterile velvet. This was used to stamp colonies onto M9 agar plates containing 1, 5, or 10 μ M proline, 0.3 M NaCl, and 0, 25, 50, 100, 250, and 500 μ M 4S-NH₂. Plates were stamped in order of increasing concentrations of 4S-NH₂, then finally onto an LB agar plate lacking 4S-NH₂. After incubation at 37°C, scanned images of replica plates were obtained and processed with ImageJ to measure growth inhibition.

5.6.15 Identification of proline-free carrier proteins

We reasoned that the Protein Data Bank may be a good resource to identify soluble proteins that express well in *E. coli*. We identified 2057 unique PDBid's corresponding to protein entities that do not contain proline residues. We limited this list to proteins that were expressed in *E. coli*, and those that existed as monomers. We confirmed that proteins were

purified from the soluble fraction, and avoided membrane proteins, enzymes, DNA binding proteins, and those that bound cofactors. One publication⁴⁴ designed many proteins that fit our criteria; in this case, two proteins from this manuscript were chosen. Codon optimization occasionally failed; in these cases, proteins were either omitted (4H3L 3), or truncated (Tako8). The final list of proline-free proteins tested here is displayed in Table 5.S2.

5.6.16 Microscopy

Images were obtained at the Biological Imaging Facility at Caltech on a Zeiss LSM800 confocal microscope. A 488 nm laser was used for fluorescence images.

5.7 Supplementary figures and tables

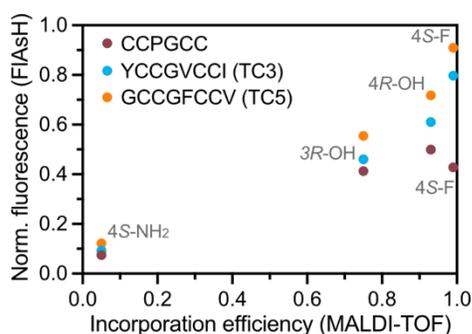


Figure 5.S1. Proline-containing TC motifs do not accurately measure proline analog incorporation. *E. coli* strain CAG18515 overexpressing the ProRS was treated with the indicated ncPro after a medium shift and F/AsH labeling. Expression of Top7, which was translationally fused to a proline-containing linker and the indicated TC tag, was then induced, and fluorescence was determined by flow cytometry. Normalized fluorescence is relative to the –Proline and +Proline controls for each strain. Compared to proline-free TC tags TC3 and TC5, normalized fluorescence for ncPro-treated samples with the canonical CCPGCC TC tag does not correlate well with known incorporation efficiencies as determined by MALDI-TOF.

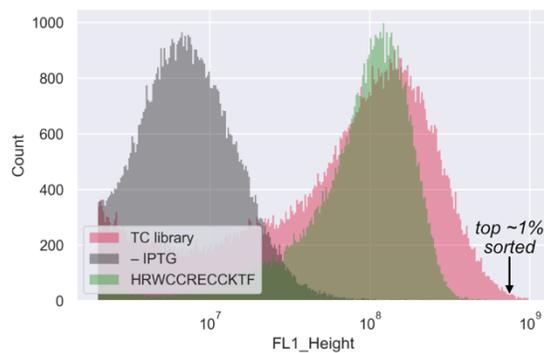


Figure 5.S2. Development of a proline-free TC motif. A library of proline-free TC motifs was fused to the C-terminus of Top7 and a proline-containing linker. Shown here is the fluorescence histogram of TC library against the previously reported proline-free TC motif used in mammalian cells: HRWCCRECKTF.⁵¹ We sorted the brightest ~1% of events once, which yielded TC3 (YCCGVCCI) and TC5 (GCCGFCCV) as the top-performing TC tags (data not shown).

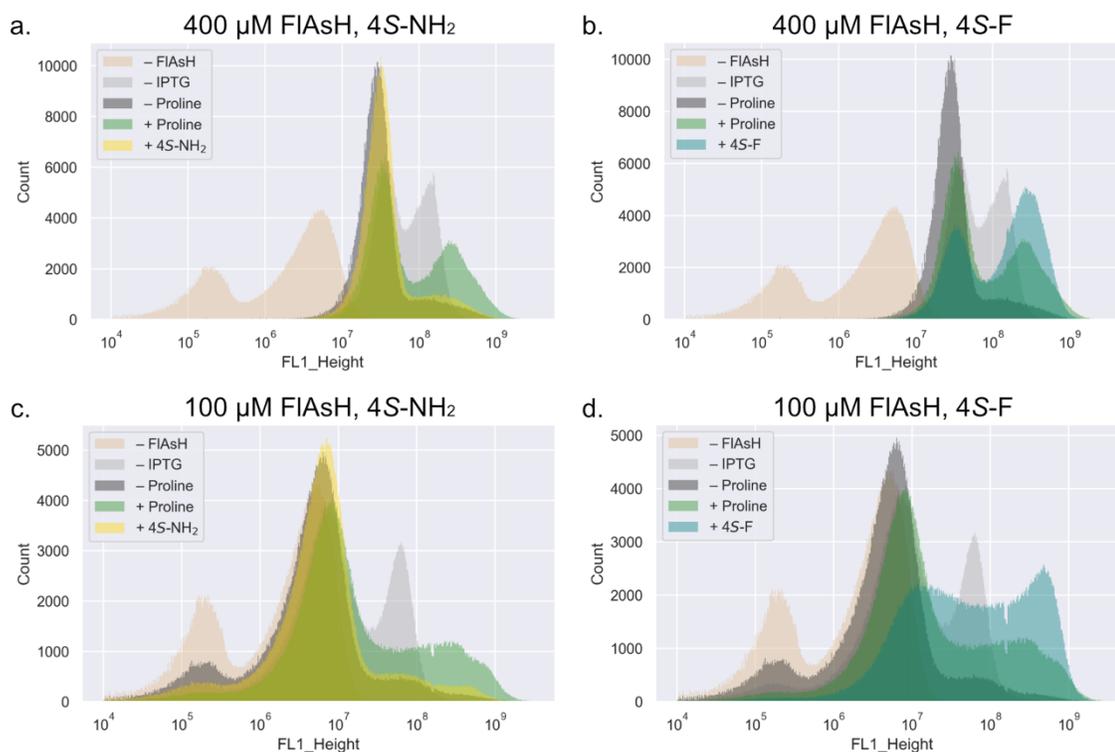


Figure 5.S3. FIASH labeling after Top7 expression leads to heterogeneous fluorescence.

Cultures were treated with the indicated concentrations of FIASH, and four-fold concentrations of EDT, for 30 min after Top7-TC3 expression under ncPro incorporation conditions. Similar histograms were obtained for a broad range of ncPro residues tested (data not shown); in these cases, the brightest ncPro population corresponded to its known incorporation efficiency. We note the two populations across all samples, and conclude that under these conditions, only a small subset of cells is sufficiently labeled with FIASH. We hypothesize that the limiting factor is FIASH entry into cells, since extended incubation of FIASH with these strains leads to a higher proportion of bright cells (i.e., Figure 5.14f).

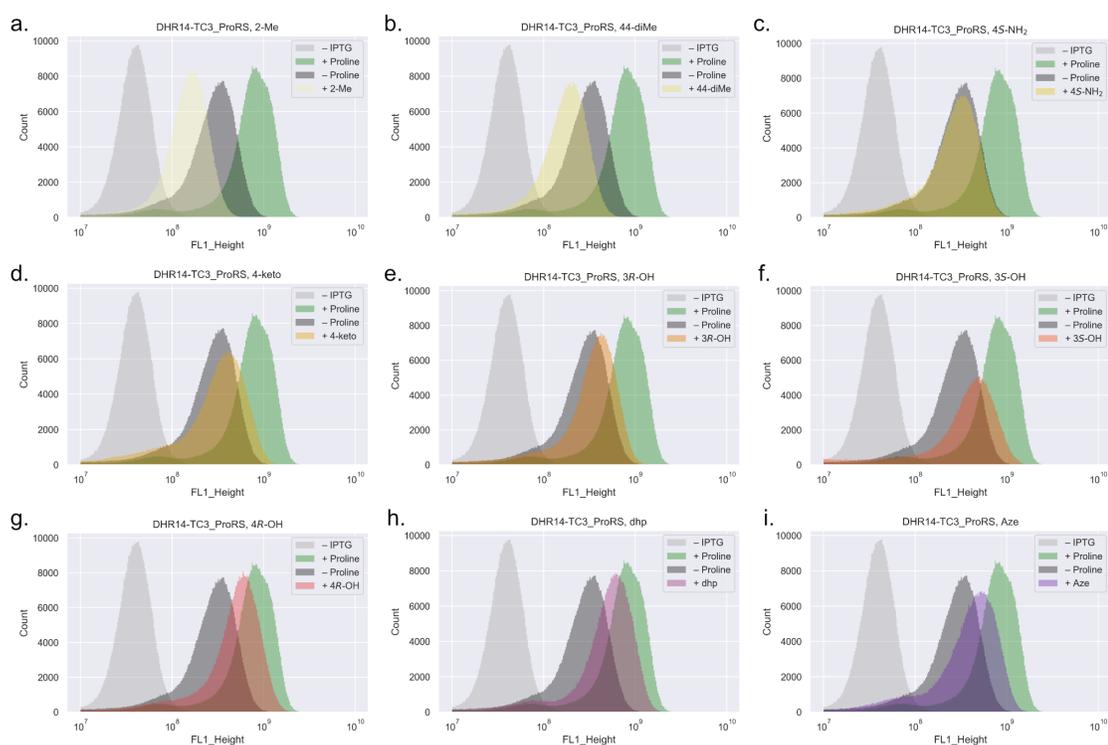


Figure 5.S4. FACS labeling after DHR14-TC3 expression can measure proline analog incorporation. Interestingly, the fluorescence data here suggest that ncPro residues 2-methylproline (2-Me) and 4,4-dimethylproline (44-diMe) inhibit residual proline uptake or incorporation, since samples treated with these proline analogs were less fluorescent than the – Proline negative control. These results are consistent with SDS-PAGE and MALDI-TOF data (Figure 2.S1, Table 2.S2). A summary of these flow cytometry histograms is displayed in Figure 5.17a.

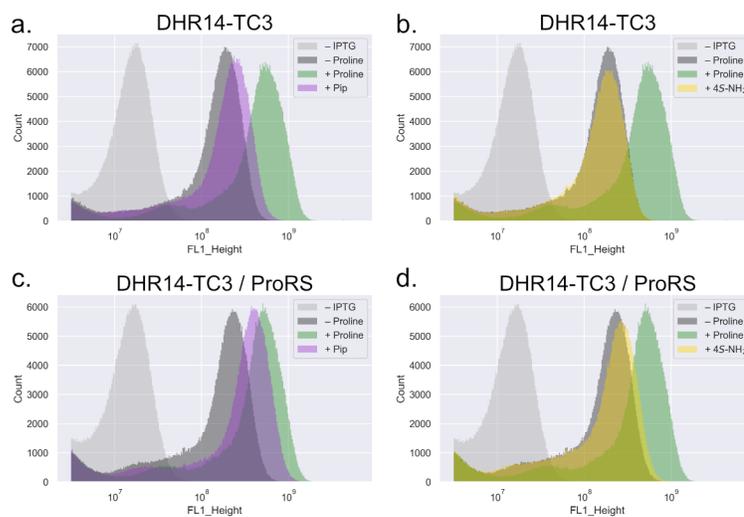


Figure 5.S5. ProRS overexpression increases proline analog incorporation. Strains harboring a plasmid either without (a-b) or with (c-d) *proS* overexpression were treated with the ncPro residues Pip (a,c) or 4S-NH₂ (b,d). A summary of these data is displayed in Figure 5.17b.

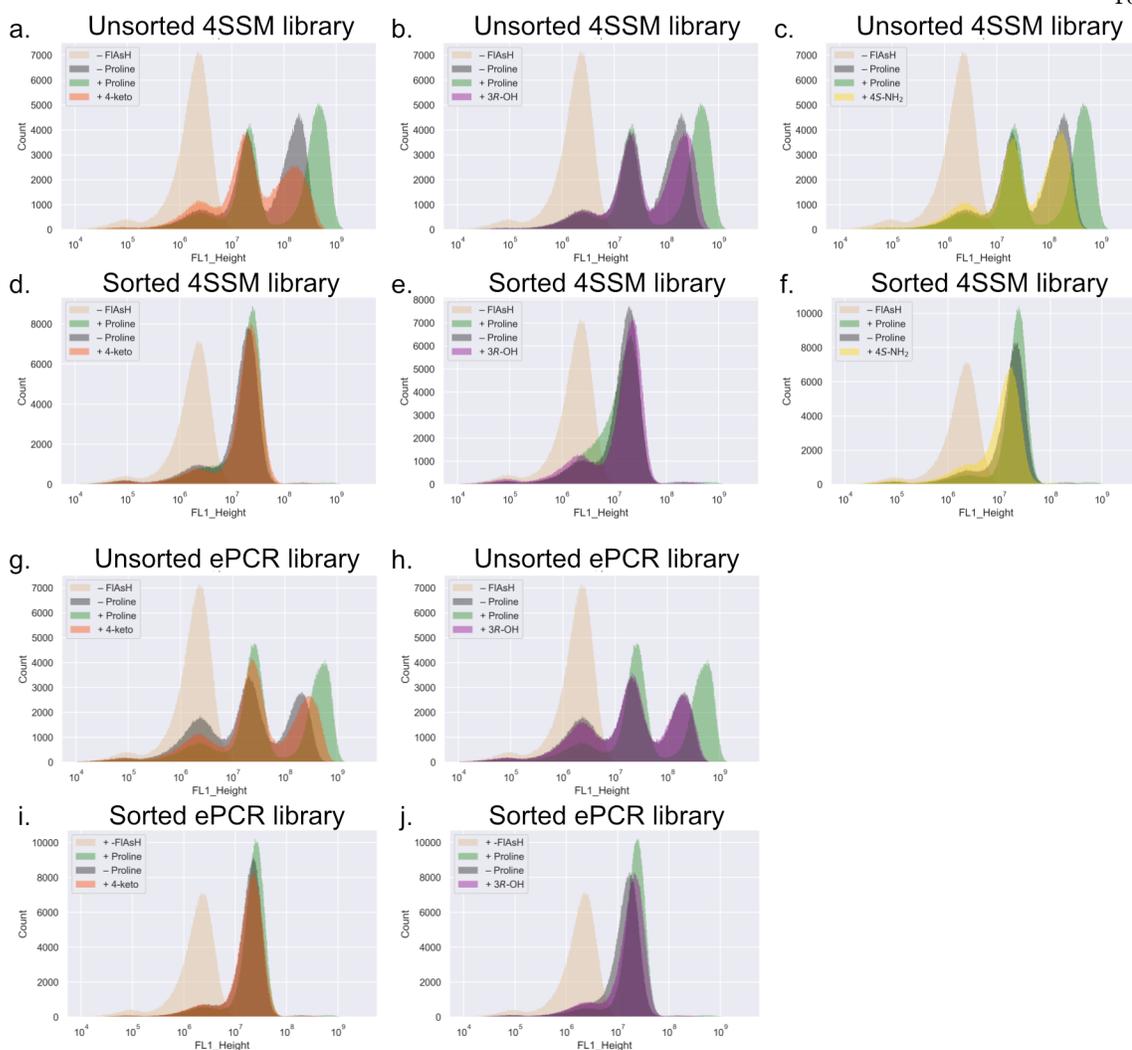


Figure 5.S6. Cheaters overwhelm library samples after sorting. 4-codon site saturation mutagenesis (4SSM, a-c) and error-prone PCR (ePCR, g-h) libraries were constructed by a restriction enzyme approach and treated with the proline analogs 4-keto (a,g), 3R-OH (b,h), and 4S-NH₂ (c). The brightest ~1% of each of these ncPro samples was sorted. Plasmid DNA was purified from sorted cells, and used to transform fresh stocks of CAG18515 for a second round of analysis by flow cytometry (d-f, i-j). As with the ePCR library treated with 4S-NH₂ shown in Figure 5.18a-b, each of the sorted samples contains only a small (~1%) population of bright cells.

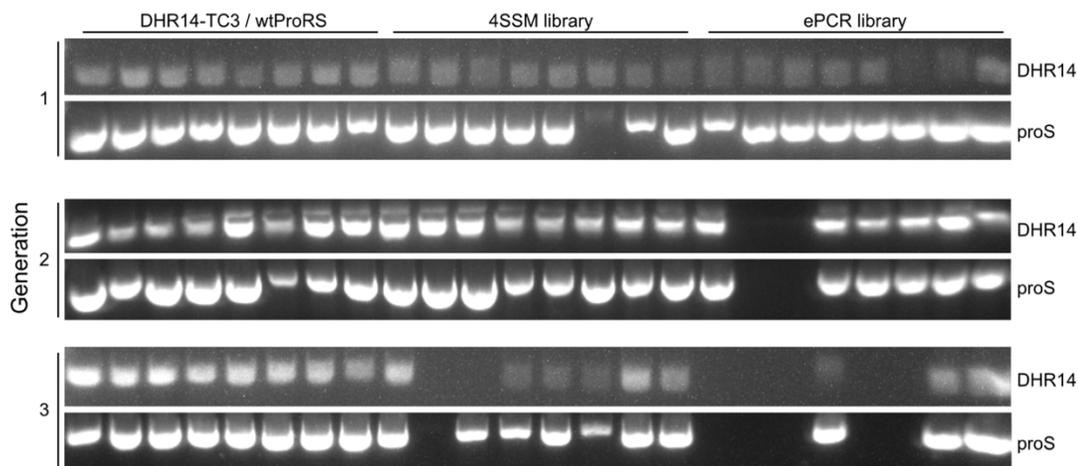


Figure 5.S7. Plasmid stability over rounds of transformation and growth. The indicated plasmid DNA was used to transform strain CAG18515, a portion of which were plated onto selective (ampicillin) agar plates. Colony PCR was used to interrogate the presence of the plasmid-borne DHR14-TC3 gene, or overexpressed *proS* gene, using primers SLB2429_analysis_fwd / SLB2429_analysis_rev, and SLB2406_fwd / SLB2406_rev, respectively. The remaining culture after transformation was grown to stationary phase in rich medium, after which plasmids were purified and used to transform subsequent generations. As opposed to previous experiments in which a restriction enzyme approach was used, the libraries here was constructed with Gibson Assembly. A summary of these data is displayed in Figure 5.818d; a colony was considered “PCR-positive” if a band was observed for both DHR14-TC3 and *proS* genes.

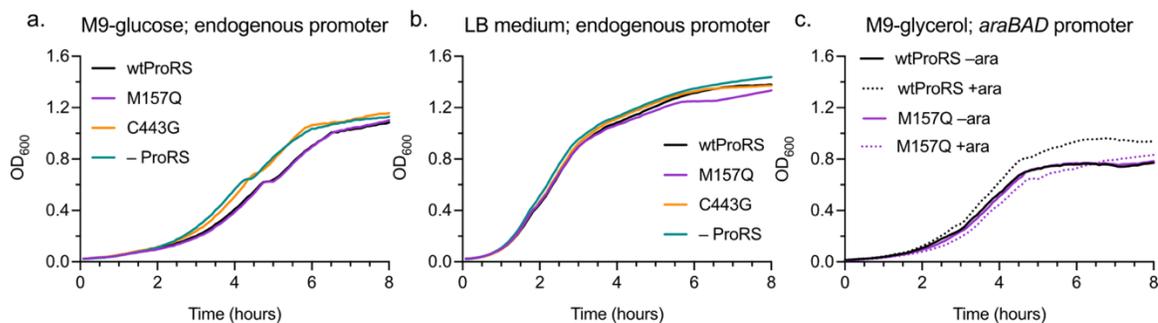


Figure 5.S8. Growth depends upon ProRS variant and promoter. a-b. *E. coli* strain CAG18515 harboring plasmid pQE80_DHR14-TC3_proS (or pQE80_DHR14-TC3) were grown in a plate reader at 37°C with shaking in 20aa M9 medium with glucose as a carbon source (a), or in LB medium (b); growth was monitored over time by tracking absorbance at 600 nm. c. CAG18515 harboring plasmid pQE80_DHR14-TC3_proS-araC, or the M157Q point mutant of *proS*, were grown similarly in M9 medium with glycerol as the carbon source; *proS* expression was induced by the addition of arabinose. Shown here is the average of three technical replicates.

Table 5.S1. Mutations rates for error-prone PCR library generation

[MnCl₂] (μM)	Mutations per <i>proS</i> gene
0	0.67 ± 1.50
200	2.13 ± 1.87
300	3.63 ± 2.12
400	7.37 ± 4.40

Table 5.S2. Amino acid sequences of proline-free proteins

Name	Amino acid sequence	MW (kDa)
Tako8	mrgshhhhhhMGQSLRTLQGHQSAVTSLQFNDNIVVSGSDD STVKVWDIKTGQSLRTLQGHQSAVTSLQFNDNIVVSGSD DSTVKVWDIKTGGsggspggspggspggsggsyccgvcci	12.4
Ika4	mrgshhhhhhMGQELVSLEGHQSAITALAFSKNIVVSGAAD GTIKVWDILTGQLLRDHDGHQSEVTALQFKDNIVVSGA KDGTVKVWYIGTGQELVSLEGHQSAITALAFSKNIVVSG AADGTIKVWDILTGQLLRDHDGHQSEVTALQFKDNIVV SGAKDGTVKVWYIGTGGsggspggspggspggsggsyccgvcci	20.7
DHR10	mrgshhhhhhMSSEKEELRERLVKIVVENAKRKGDDTEEAR EAAREAFELVREAAERAGIDSSEVLELAIRLIKEVVENAQ REGYDISEAARAAAEAFKRVAEAAK RAGITSSEVLELAI RLIKEVVENAQREGYDISEAARAAAEAFKRVAEAAKRA GITSSETLKRAIEEIRKRVEEAQREGNDISEAARQAEEFR KKAELKRRGDGWLEggspggspggspggsggsyccgvcci	26.2
DHR14	mrgshhhhhhMDSEEVNERVKQLAEKAKEATDKEEVIEIVK ELAE LAKQSTDSELVNEIVKQLAEVAKEATDKELVIYIV KILAE LAKQSTDSELVNEIVKQLAEVAKEATDKELVIYIV KILAE LAKQSTDSELVNEIVKQLEEVAKEATDKELVEHIE KILEELKKQSTDGWLEggspggspggspggsggsyccgvcci	22.1
3H5L_2	mrgshhhhhhNEDDMKKLYKQMVQELEKARDRMEKLYKE MVELIQKAIELMRKIFQEVKQVEVEKAIEEMKKLYDEAKK KIEQMIQQIKQGGDKQKMEELLKRAKEEMKKVKDKME KLEELKQIMQEAKQKMEKLLKQLKEEMKKMKEKMEK LLKEMKQRMEEVKKKMDGDDELLEKIKKNIDDLKIAE DLIKKAENIKEAKKIAEQLVKRAKQLIEKAKQVAEELIK KILQLIEKAKEIAEKVLKGLEggspggspggspggsggsyccgvcci	32.4
Blo t 21	mrgshhhhhhNTATQRFHEIEKFLHITHEVDDLEKTGNKDE KARLLRELVSEAFIEGSRGYFQRELKRTDLDLLEKFNFE AALATGDLLLKDLKALQKRVQDSEggspggsp	14.6
Pkd2	mrgshhhhhhMGSTAIGINDTYSEVKSDLAQQKAEMELSDLI RKG YHKALVKLKLKNTVDDISESLRQGGGKLNFDDEL QDLKGGKHTDAEIEAIFTKYDQDGDQELTEHEHQMRD DLEKEREDLDLDHSSLggspggspggspggsggsyccgvcci	17.5
Utr-SR1	mrgshhhhhhDMDLDSYQIALEEVLTWLLSAEDTFQEQQDDI SDDVEDVKEQFATHETFMMELSAHQSSVGSVLQAGNQL MTQGTLSDDEEFIEIQEMTLLNARWEALRVESMERQSR LHDALMELQKKQLQQLggspggspggspggsggsyccgvcci	17.7
Top7	MGDIQVQVNIDDNGKNFDYTYTVTTESELQKVLNELKD YIEEQGAKRARISITARTEKEAEKFAAILIKVFAELGYNDI NVTWDGDTVTVEGQLEGGSLepggspggspggsggsyccgvcci	13.1

*The N-terminal 6xHis tag and C-terminal proline linker + TC3 sequences are in lowercase; the sequences corresponding to the proline-free protein are in UPPERCASE.

5.8 References

- (1) Ling, J.; Reynolds, N.; Ibba, M. Aminoacyl-tRNA Synthesis and Translational Quality Control. *Annu Rev Microbiol* **2009**, *63*, 61–78. <https://doi.org/10.1146/annurev.micro.091208.073210>.
- (2) Chin, J. W. Expanding and Reprogramming the Genetic Code. *Nature* **2017**, *550*, 53–60. <https://doi.org/10.1038/nature24031>.
- (3) Yoo, T. H.; Tirrell, D. A. High-Throughput Screening for Methionyl-tRNA Synthetases That Enable Residue-Specific Incorporation of Noncanonical Amino Acids into Recombinant Proteins in Bacterial Cells. *Angew Chem Int Ed* **2007**, *46*, 5340–5343. <https://doi.org/10.1002/anie.200700779>.
- (4) Ngo, J. T.; Champion, J. A.; Mahdavi, A.; Tanrikulu, I. C.; Beatty, K. E.; Connor, R. E.; Yoo, T. H.; Dieterich, D. C.; Schuman, E. M.; Tirrell, D. A. Cell-Selective Metabolic Labeling of Proteins. *Nat Chem Biol* **2009**, *5* (10), 715–717. <https://doi.org/10.1038/nchembio.200>.
- (5) Renner, C.; Alefelder, S.; Bae, J. H.; Budisa, N.; Huber, R.; Moroder, L. Fluoroprolines as Tools for Protein Design and Engineering. *Angew Chem Int Ed* **2001**, *40* (5), 923–925. [https://doi.org/10.1002/1521-3773\(20010302\)40:5<923::AID-ANIE923>3.0.CO;2-#](https://doi.org/10.1002/1521-3773(20010302)40:5<923::AID-ANIE923>3.0.CO;2-#).
- (6) Breunig, S. L.; Tirrell, D. A. Incorporation of Proline Analogs into Recombinant Proteins Expressed in *Escherichia coli*. *Methods Enzymol* **2021**, *656*, 545–571. <https://doi.org/10.1016/BS.MIE.2021.05.008>.
- (7) Kim, W.; George, A.; Evans, M.; Conticello, V. P. Cotranslational Incorporation of a Structurally Diverse Series of Proline Analogues in an *Escherichia coli* Expression System. *ChemBioChem* **2004**, *5* (7), 928–936. <https://doi.org/10.1002/cbic.200400052>.
- (8) Packer, M. S.; Liu, D. R. Methods for the Directed Evolution of Proteins. *Nat Rev Genet* **2015**, *16*, 379–394. <https://doi.org/10.1038/nrg3927>.
- (9) Krahn, N.; Tharp, J. M.; Crnković, A.; Söll, D. Engineering Aminoacyl-tRNA Synthetases for Use in Synthetic Biology. *Enzymes* **2020**, *48*, 351. <https://doi.org/10.1016/BS.ENZ.2020.06.004>.
- (10) Chatterjee, A.; Xiao, H.; Schultz, P. G. Evolution of Multiple, Mutually Orthogonal Prolyl-tRNA Synthetase/tRNA Pairs for Unnatural Amino Acid Mutagenesis in *Escherichia coli*. *Proc Natl Acad Sci* **2012**, *109* (37), 14841–14846. <https://doi.org/10.1073/pnas.1212454109>.
- (11) Carpenter, H. E. Biomaterials Design: Creation of a Novel Orthogonal Translational System in *E. coli* for the Site-Specific Incorporation of Proline Analogues, Emory University, 2002.
- (12) Johnson, J. A.; Lu, Y. Y.; Van Deventer, J. A.; Tirrell, D. A. Residue-Specific Incorporation of Non-Canonical Amino Acids into Proteins: Recent Developments and Applications. *Curr Opin Chem Biol* **2010**, *14*, 774–780. <https://doi.org/10.1016/j.cbpa.2010.09.013>.

- (13) Steiner, T.; Hess, P.; Bae, J. H.; Wiltschi, B.; Moroder, L.; Budisa, N. Synthetic Biology of Proteins: Tuning GFPs Folding and Stability with Fluoroproline. *PLoS One* **2008**, *3* (2), e1680. <https://doi.org/10.1371/journal.pone.0001680>.
- (14) Deepankumar, K.; Nadarajan, S. P.; Ayyadurai, N.; Yun, H. Enhancing the Biophysical Properties of mRFP1 through Incorporation of Fluoroproline. *Biochem Biophys Res Commun* **2013**, *440* (4), 509–514. <https://doi.org/10.1016/j.bbrc.2013.09.062>.
- (15) To, T. M. T.; Kubyshkin, V.; Schmitt, F. J.; Budisa, N.; Friedrich, T. Residue-Specific Exchange of Proline by Proline Analogs in Fluorescent Proteins: How “Molecular Surgery” of the Backbone Affects Folding and Stability. *J Vis Exp* **2022**, *2022* (180). <https://doi.org/10.3791/63320>.
- (16) Shoulders, M. D.; Raines, R. T. Collagen Structure and Stability. *Annu Rev Biochem* **2009**, *78*, 929–958. <https://doi.org/10.1146/annurev.biochem.77.032207.120833>.
- (17) Ganguly, H. K.; Basu, G. Conformational Landscape of Substituted Prolines. *Biophys Rev* **2020**, *12*, 25–39. <https://doi.org/10.1007/s12551-020-00621-8>.
- (18) Ong, W. J. H.; Alvarez, S.; Leroux, I. E.; Shahid, R. S.; Samma, A. A.; Peshkepaja, P.; Morgan, A. L.; Mulcahy, S.; Zimmer, M. Function and Structure of GFP-like Proteins in the Protein Data Bank. *Mol Biosyst* **2011**, *7* (4), 984–992. <https://doi.org/10.1039/C1MB05012E>.
- (19) Fang, K. Y. Modulating Biophysical Properties of Insulin with Non-Canonical Mutagenesis at Position B28, California Institute of Technology, 2017. <https://doi.org/10.7907/Z9V40S6J>.
- (20) Cabantous, S.; Terwilliger, T. C.; Waldo, G. S. Protein Tagging and Detection with Engineered Self-Assembling Fragments of Green Fluorescent Protein. *Nat Biotechnol* **2005**, *23* (1), 102–107. <https://doi.org/10.1038/nbt1044>.
- (21) de Boer, H. A.; Comstock, L. J.; Vasser, M. The Tac Promoter: A Functional Hybrid Derived from the Trp and Lac Promoters. *Proc Natl Acad Sci* **1983**, *80* (1), 21–25. <https://doi.org/10.1073/pnas.80.1.21>.
- (22) Deng, A.; Boxer, S. G. Structural Insight into the Photochemistry of Split Green Fluorescent Proteins: A Unique Role for a His-Tag. *J Am Chem Soc* **2018**, *140*, 375–381. <https://doi.org/10.1021/jacs.7b10680>.
- (23) Guzman, L.-M.; Belin, D.; Carson, M. J.; Beckwith, J. Tight Regulation, Modulation, and High-Level Expression by Vectors Containing the Arabinose PBAD Promoter. *J Bacteriol* **1995**, *177* (14), 4121–4130. <https://doi.org/10.1128/jb.177.14.4121-4130.1995>.
- (24) Kamiyama, D.; Sekine, S.; Barsi-Rhyne, B.; Hu, J.; Chen, B.; Gilbert, L. A.; Ishikawa, H.; Leonetti, M. D.; Marshall, W. F.; Weissman, J. S.; Huang, B. Versatile Protein Tagging in Cells with Split Fluorescent Protein. *Nat Commun* **2016**, *7*, 11046. <https://doi.org/10.1038/ncomms11046>.

- (25) Cabantous, S.; Waldo, G. S. In Vivo and in Vitro Protein Solubility Assays Using Split GFP. *Nat Methods* **2006**, *3* (10), 845–854. <https://doi.org/10.1038/nmeth932>.
- (26) Feinberg, E. H.; VanHoven, M. K.; Bendesky, A.; Wang, G.; Fetter, R. D.; Shen, K.; Bargmann, C. I. GFP Reconstitution Across Synaptic Partners (GRASP) Defines Cell Contacts and Synapses in Living Nervous Systems. *Neuron* **2008**, *57* (3), 353–363. <https://doi.org/10.1016/j.neuron.2007.11.030>.
- (27) Pinaud, F.; Dahan, M. Targeting and Imaging Single Biomolecules in Living Cells by Complementation-Activated Light Microscopy with Split-Fluorescent Proteins. *Proc Natl Acad Sci* **2011**, *108* (24), E201–E210. <https://doi.org/10.1073/pnas.1101929108>.
- (28) Hyun, S. I.; Maruri-Avidal, L.; Moss, B. Topology of Endoplasmic Reticulum-Associated Cellular and Viral Proteins Determined with Split-GFP. *Traffic* **2015**, *16* (7), 787–795. <https://doi.org/10.1111/tra.12281>.
- (29) van Engelenburg, S. B.; Palmer, A. E. Imaging Type-III Secretion Reveals Dynamics and Spatial Segregation of Salmonella Effectors. *Nat Methods* **2010**, *7*:4 **2010**, *7* (4), 325–330. <https://doi.org/10.1038/nmeth.1437>.
- (30) Kaddoum, L.; Magdeleine, E.; Waldo, G. S.; Joly, E.; Cabantous, S. One-Step Split GFP Staining for Sensitive Protein Detection and Localization in Mammalian Cells. *Biotechniques* **2010**, *49* (4), 727–736. <https://doi.org/10.2144/000113512>.
- (31) Kim, Y. E.; Kim, Y. N.; Kim, J. A.; Kim, H. M.; Jung, Y. Green Fluorescent Protein Nanopolygons as Monodisperse Supramolecular Assemblies of Functional Proteins with Defined Valency. *Nat Commun* **2015**, *6*:1 **2015**, *6* (1), 1–9. <https://doi.org/10.1038/ncomms8134>.
- (32) Lawrence, M. S.; Phillips, K. J.; Liu, D. R. Supercharging Proteins Can Impart Unusual Resilience. *J Am Chem Soc* **2007**, *129*, 10110–10112. <https://doi.org/10.1021/ja071641y>.
- (33) Huang, Y.-M.; Nayak, S.; Bystroff, C. Quantitative in Vivo Solubility and Reconstitution of Truncated Circular Permutants of Green Fluorescent Protein. *Protein Sci* **2011**, *20*, 1775–1780. <https://doi.org/10.1002/pro.735>.
- (34) Minks, C.; Alefelder, S.; Moroder, L.; Huber, R.; Budisa, N. Towards New Protein Engineering: In Vivo Building and Folding of Protein Shuttles for Drug Delivery and Targeting by the Selective Pressure Incorporation (SPI) Method. *Tetrahedron* **2000**, *56* (48), 9431–9442. [https://doi.org/10.1016/S0040-4020\(00\)00827-9](https://doi.org/10.1016/S0040-4020(00)00827-9).
- (35) Schellenberger, V.; Wang, C.-W.; Geething, N. C.; Spink, B. J.; Campbell, A.; To, W.; Scholle, M. D.; Yin, Y.; Yao, Y.; Bogin, O.; Cleland, J. L.; Silverman, J.; Stemmer, W. P. C. A Recombinant Polypeptide Extends the in Vivo Half-Life of Peptides and Proteins in a Tunable Manner. *Nat Biotechnol* **2009**, *27* (12), 1186–1190. <https://doi.org/10.1038/nbt.1588>.

- (36) Wittrup, K. D.; Mann, M. B.; Fenton, D. M.; Tsai, L. B.; Bailey, J. E. Single-Cell Light Scatter as a Probe of Refractile Body Formation in Recombinant *Escherichia coli*. *Bio/Technology*. 1988, pp 423–426. <https://doi.org/10.1038/nbt0488-423>.
- (37) Lavis, L. D.; Raines, R. T. Bright Ideas for Chemical Biology. *ACS Chem Biol* **2008**, *3* (3), 142–155. <https://doi.org/10.1021/cb700248m>.
- (38) Griffin, B. A.; Adams, S. R.; Tsien, R. Y. Specific Covalent Labeling of Recombinant Protein Molecules Inside Live Cells. *Science (1979)* **1998**, *281*, 269–272. <https://doi.org/10.1126/science.281.5374.269>.
- (39) Landgraf, D.; Okumus, B.; Chien, P.; Baker, T. A.; Paulsson, J. Segregation of Molecules at Cell Division Reveals Native Protein Localization. *Nat Methods* **2012**, *9* (5), 480–482. <https://doi.org/10.1038/nmeth.1955>.
- (40) Stroffekova, K.; Proenza, C.; Beam, K. G. The Protein-Labeling Reagent FIAsh-EDT₂ Binds Not Only to CCXXCC Motifs but Also Non-Specifically to Endogenous Cysteine-Rich Proteins. *Pflugers Arch* **2001**, *442* (6), 859–866. <https://doi.org/10.1007/s004240100619>.
- (41) Dantas, G.; Watters, A. L.; Lunde, B. M.; Eletr, Z. M.; Isern, N. G.; Roseman, T.; Lipfert, J.; Doniach, S.; Tompa, M.; Kuhlman, B.; Stoddard, B. L.; Varani, G.; Baker, D. Mis-Translation of a Computationally Designed Protein Yields an Exceptionally Stable Homodimer: Implications for Protein Engineering and Evolution. *J Mol Biol* **2006**, *362* (5), 1004–1024. <https://doi.org/10.1016/j.jmb.2006.07.092>.
- (42) Adams, S. R.; Campbell, R. E.; Gross, L. A.; Martin, B. R.; Walkup, G. K.; Yao, Y.; Llopis, J.; Tsien, R. Y. New Biarsenical Ligands and Tetracysteine Motifs for Protein Labeling in Vitro and in Vivo: Synthesis and Biological Applications. *J Am Chem Soc* **2002**, *124* (21), 6063–6076. <https://doi.org/10.1021/ja017687n>.
- (43) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res* **2018**, *46* (W1), W296–W303. <https://doi.org/10.1093/nar/gky427>.
- (44) Brunette, T.; Parmeggiani, F.; Huang, P.-S.; Bhabha, G.; Ekiert, D. C.; Tsutakawa, S. E.; Hura, G. L.; Tainer, J. A.; Baker, D. Exploring the Repeat Protein Universe through Computational Protein Design. *Nature* **2015**, *528*, 580–584. <https://doi.org/10.1038/nature16162>.
- (45) Navaratna, T.; Atangcho, L.; Mahajan, M.; Subramanian, V.; Case, M.; Min, A.; Tresnak, D.; Thurber, G. M. Directed Evolution Using Stabilized Bacterial Peptide Display. *J Am Chem Soc* **2020**, *142*, 1882–1894. <https://doi.org/10.1021/jacs.9b10716>.
- (46) Ramesh, B.; Frei, C. S.; Cirino, P. C.; Varadarajan, N. Functional Enrichment by Direct Plasmid Recovery after Fluorescence Activated Cell Sorting. *Biotechniques* **2015**, *59* (3), 157–161. <https://doi.org/10.2144/000114329>.

- (47) Tam, L. M.; Price, N. E.; Wang, Y. Molecular Mechanisms of Arsenic-Induced Disruption of DNA Repair. *Chem Res Toxicol* **2020**, *33* (3), 709–726. <https://doi.org/10.1021/acs.chemrestox.9b00464>.
- (48) Kanagawa, T. Bias and Artifacts in Multitemplate Polymerase Chain Reactions (PCR). *J Biosci Bioeng* **2003**, *96* (4), 317–323. [https://doi.org/10.1016/s1389-1723\(03\)90130-7](https://doi.org/10.1016/s1389-1723(03)90130-7).
- (49) Datsenko, K. A.; Wanner, B. L. One-Step Inactivation of Chromosomal Genes in *Escherichia coli* K-12 Using PCR Products. *Proc Natl Acad Sci* **2000**, *97* (12), 6640–6645. <https://doi.org/10.1073/pnas.120163297>.
- (50) Yuet, K. P. Tools For Spatiotemporally Specific Proteomic Analysis In Multicellular Organisms By, California Institute of Technology, 2016. <https://doi.org/10.7907/z9vd6wdh>.
- (51) Martin, B. R.; Giepmans, B. N.; Adams, S. R.; Tsien, R. Y. Mammalian Cell-Based Optimization of the Biarsenical- Binding Tetracysteine Motif for Improved Fluorescence and Affinity. *Nat Biotechnol* **2005**, *23* (10), 1308–1314. <https://doi.org/10.1038/nbt1136>.