Modulating Biophysical Properties of Insulin with Non-Canonical Mutagenesis at Position B28

Thesis by
Katharine Y. Fang

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

California Institute of Technology
Pasadena, California, USA

2017

(Defended January 17, 2017)

For my family.

ACKNOWLEDGEMENTS

To my advisor, Professor David A. Tirrell: Thank you for your support and guidance, and especially for taking the chance on a naïve first-year who has required quite a bit of patience over the years. Over these past years, with your help, I have transitioned to a scientist who feels confident and capable of facing whatever challenges are ahead.

To my committee members, Professors H. Teresa Ku, Stephen Mayo, and Zhen-Gang Wang: Thank you for your advice and time.

To my mentor outside of the lab, Dr. Mike Vicic: Thank you for all the life and work advice you have given me over the years. I will always be grateful for having had the opportunity to work alongside you and teach ChE130.

To my long-suffering collaborator, lab mate, and friend, Seth Lieblich: Thank you for working with me. This project and thesis were only possible because of the discussions we had, the encouragement you provided, and our dynamic as a team.

To my mentor, Dr. Beverly Lu, and my mentee, Mary Boyajian: Thank you for mentoring me, teaching me how to mentor, and giving me the opportunity to serve as a mentor.

To all my colleagues, staff, and lab mates: Thank you for the all the useful suggestions, help, and discussions over the years.

To my family: Thank you for your unconditional love and support, and everything else in-between.

ABSTRACT

Non-canonical amino acids are tools for altering the chemical and physical properties of proteins, providing a facile strategy to engineer proteins with novel properties, especially where canonical amino acid mutagenesis has exhausted nearly all avenues for further optimization. Insulin, for example, is one of the most widely studied therapeutic proteins; however, non-canonical insulin engineering is a subfield that has been largely unexplored. To this end, my thesis research has focused on the use of non-canonical amino acids to understand and engineer the biophysical properties of insulin.

Protein structure and function are sensitive to the smallest of changes; even small differences such as a single atom substitution or change in stereo-orientation can cause large, unsuspected, and unpredictable global changes. Chapters 2 to 4 describe substitutions at the 4th position of proline at position 28 of insulin's B chain (ProB28) in a manner analogous to the structure-activity-relationships that are widely used in medicinal chemistry. Canonical mutagenesis at ProB28 has led to the discovery of rapid-acting insulins (RAIs): a class of therapeutic insulins with enhanced pharmacokinetic properties. Therefore, we chose to incorporate proline analogs with substitutions such as hydroxyl and fluoro groups, and different ring compositions to assess their effects on the biophysical properties (i.e. stability, dissociation rates and oligomerizations states) of insulin. Chapter 2 describes the discovery of a hydroxyinsulin variant with faster hexamer dissociation rates and enhanced stability compared to wild-type insulin *in vitro*. We find crystallographic evidence of a novel hydrogen bond in the insulin dimer interface which we hypothesize stabilizes the

insulin dimer state. To complement the findings in chapter 2, chapters 3 and 4 detail an investigation of the importance of hydrophobic and nonpolar interactions for modulating the biophysical properties of insulin.

Establishing structure-activity-relationships for insulin will create new opportunities for further engineering using non-canonical amino acids. In chapter 5, we describe progress towards a general, simple screening method to discover new aminoacyl-tRNA synthetases for the incorporation of non-canonical amino acids in *E. coli*. Implementing such a high-throughput screening system will allow scientists to perform medicinal chemistry on proteins and discover new or improved therapeutics to help manage human diseases.

TABLE OF CONTENTS

LIST OF FIGURES & TABLES

Figures

Tables

Chapter 1 – Introduction

Importance of biological polymers. Life is encoded in biological polymers and is summarized by the central dogma of biology: DNA is first transcribed into (messenger) RNA, which is then translated into proteins. Often, human diseases arise from changes in the protein's amino acid sequence, which can be caused[1] by a genetic mutation at the chromosomal level, an error in transcription by the RNA polymerase, or misacylation by an aminoacyl tRNA synthetase during translation by the ribosome. The accumulation of mutations within a gene and their corresponding structural changes within a protein[2,3] can have extreme consequences, especially if the protein is involved in an important signaling pathway. For example, one of the causes of Alzheimer's disease[4] (AD) has been speculated to be due to one mutation (A42T) in the amyloid precursor protein (APP), which renders its downstream protein, amyloid-β (Aβ), more prone to form and accumulate fibrillar aggregates. In addition to mutations in a protein's primary structure, absence of critical proteins can also have dire consequences. This is highlighted by *diabetes mellitus*, a chronic pancreatic illness[5], which is caused by either the body's inability to produce (Type I) or respond to (Type II) the blood glucose regulating protein, insulin. Since many diseases arise due to specific protein irregularities, studying and utilizing these complex biological polymers can offer beneficial alternative treatment methods.

Protein therapeutics. One advantage for the clinical use of protein therapeutics over small molecule drugs is the specificity of the protein to induce a complex set of functions[6]. Proteins have naturally evolved from the 20 canonical amino acids (cAAs) whose unique side-chains participates in determining the structure of the macromolecule, which in turn determines its function. Since the accessible sequence space for proteins is near limitless,

the potential to create is enormous. No synthetically derived non-proteinaceous molecule can replicate a protein's shape, specificity, and therefore, is able to function at the same molecular level. In the field of protein therapeutics, the relationship between structure and function is navigated to produce proteins that can replace or improve upon the effects of existing (or deficient) cellular functions with the goal of combating illness. If a disease is caused by the absence of a protein, then treatment should naturally involve providing the patient with an exogenous supply. This was the case with the first therapeutic protein, insulin, which was described as an endocrine hormone in the late 19th century[7], and to date no small-molecule or other alternative has been discovered that can interact with the insulin receptor in the presence of glucose in the same way as insulin. Subsequent work led to the use of insulin as a life-saving treatment for diabetes nearly thirty years later for which Frederick Banting and J. R. R. Macleod were awarded the Nobel prize in 1923[7]. Recombinant DNA (recDNA) technology has further facilitated the clinical use of proteins[8] and has helped the pharmaceutical industry achieve production on relevant scales[6]. For example, insulin for treatment was originally derived solely from bovine or porcine sources; it has since been replaced by recombinant insulin, which was approved by the Food and Drug Administration (FDA) in 1982[8]. Developing protein therapeutics also allows scientists to take advantage of having complete control over primary structure and post-translational modifications. Currently there is much interest in using the specificity of proteins to target specific cell types (e.g. cancer cells) and deliver a coupled small-molecule[9]. Ideally, proteins can be developed with the desired activity profile, clearance properties, and physical characteristics without interfering with other cellular processes and with limited toxicity[10].

However, there remain several drawbacks: *in vivo* limitations, such as clearance pathways through protease activity, kidney filtration and the production of neutralizing antibodies by the body's immune response[11], and *in vitro* formulation (i.e. long-term stability) challenges[12]. These shortcomings are areas of current research[13,14] and the work described in this thesis is aimed at demonstrating the potential usefulness of non-canonical amino acids (ncAAs) in the field of therapeutic protein engineering.

Therapeutic protein engineering. Research towards addressing the limitations of therapeutic proteins can be broadly categorized into two major strategies[11,12]: first, development of medical devices for greater ease of delivery (e.g. subcutaneous injection needles), and second, engineering of the protein itself (e.g. biophysical studies[15]). The latter can be further subdivided into conjugation methods and mutating the protein's amino acid sequence. Conjugation methods have largely been used for the extension of the *in vivo* half-life, which is dominated by attaching long polymers[11] (e.g. PEGylation) or fatty acids that can bind to serum proteins[16,17]. However, modifying intrinsic protein properties, such as stability and solubility[18], is not straightforward, especially when coupled with the FDA's requirement that pharmaceutical formulations of therapeutic proteins must maintain activity for at least two years[19] under proper storage conditions (i.e. 4°C). The functional state of proteins is considered metastable[20,21] and preventing the undesired (and perhaps more stable) aggregated states from forming is a current challenge. This formulation problem[22,23] is especially difficult considering the propensity of protein aggregation at the higher concentrations[24] which therapeutic formulations typically require[18,25]. For example, formulating therapeutic insulin involves the addition of zinc ions and phenolic ligands to

form its most stable, R_6 hexamer state[26]; this is done so as to prevent the formation of insulin fibrils for long-term storage. To obtain desired pharmaceutical properties, there are often tradeoffs: e.g. the stable R_6 insulin hexamer state is not bioactive, and upon injection it requires a lag time to dissociate to its active, monomeric form[27]. This lag time places on the patient the burden of monitoring blood glucose levels and injecting insulin prior to meals; it is unsurprising that patient compliance is reported to be a global issue[28] among diabetics. Thus, important therapeutic proteins such as insulin warrant further investigation to engineer a molecule which can better balance its therapeutic value with practical concerns regarding its effective use and stability. However, the primary structure of insulin is highly conserved, which places limits for the mutational analyses that can be done to improve upon its current qualities with the canonical toolbox of amino acids (cAAs).

Non-canonical amino acids. New chemical and physical properties can be introduced into the proteome through the incorporation[29-31] of ncAAs. There are three widely used approaches for the incorporation of ncAAs: residue-specific, site-specific, and through *in vitro* or synthetic means. The residue-specific approach circumvents the need for genetic manipulation by taking advantage of the promiscuity of the endogenous translational machinery. The use of auxotrophic host strains permits global replacement of a specific residue by selective pressure; this is done with the addition of the non-canonical analogue into the culture medium in place of the canonical amino acid[29]. Most site-specific methods rely on amber codon suppression and the development of exogenous components for protein synthesis; however, this method generally is limited by low protein yields due to the termination of translation in response to the stop codon[32,33]. Recent work has addressed

some of the truncation issues through genomically recoded organisms[34] (GROs); however, the fitness of these engineered strains remains far below that of wild-type strains[35] and has, thus far, only been demonstrated in *Escherichia coli*[36]. The last category for ncAA incorporation involves non-cellular means. These include cell-free protein synthesis[37] and synthetic methods (i.e. chemical acylation of ncAA onto its cognate tRNA[31], and peptide synthesis) which can be implemented either residue- or site-specifically.

Residue-specific incorporation of ncAA offers a simple approach to study proteins of interest for a variety of applications, such as proteomic analysis, live cell imaging, and producing engineered proteins at high yields. The use of ncAAs can largely be described by two broad categories: the introduction of reactive moieties allowing for downstream enrichment, quantification and identification[38], and as probes to engineer and understand mechanistic protein behavior:

Use of ncAA for proteomics. Several reactive ncAA with chemical handles have been incorporated into proteins for subsequent modification using reactions such as azide-alkyne coupling[39-41]. This method is termed biorthogonal non-canonical amino acid tagging (BONCAT)[40]; here, the orthogonal moiety is analogous to the use of radiolabeled compounds in traditional pulse-chase experiments to distinguish between different proteomic states. The reactive handle allows for enrichment of newly synthesized proteins to reduce sample complexity prior to mass spectrometry and can be combined with stable isotope labeling in cell culture (SILAC) for quantification[42,43] (e.g. up- and down-regulated proteins in response to a stimulus). BONCAT proteomics has been demonstrated in several complex biological systems[44-46].

Use of ncAA for protein engineering. The relationships among protein sequence, structure, and function are central to life, and have been undergoing natural evolution using the twenty cAAs. Through the introduction of ncAAs into proteins, the available primary sequence space and protein function space can be further expanded beyond the restrictions of cAAs. It has been widely demonstrated in the literature that incorporating ncAAs into proteins can not only give access to new chemistries, but also affect a protein's physical properties. For example, the thermostability of leucine zipper peptides has been increased by the introduction of fluorinated amino acids[47]. However, regardless of the effects, the protein engineer is limited by the cell's ability to utilize ncAAs. To this end, there is significant interest in developing techniques that allow the translational machinery to incorporate amino acids with novel side chains. This will be described further in Chapter 5 where a general, high-throughput method is given for engineering aminoacyl-tRNA synthetases to accommodate ncAAs during translation.

Proline. The cyclic structure of proline gives its unique traits: puckering of the prolyl-ring (*exo* versus *endo* conformation), *cis-trans* isomerization of the amide peptide bond (due to the smaller energy difference necessary for *cis-trans* isomerization[48] compared to the other cAAs), and constrained dihedral angles. There has been much work done to understand the role of critical prolines in proteins; some of this work involves proline analogs[49-51] to modulate the prolyl ring pucker and its propensity to undergo cis-trans isomerization by altering the substituent group on the prolyl ring[52]. For example, non-canonical prolines (ncPro), specifically hydroxy-prolines and fluoroprolines, have contributed to understanding

collagen structure and stability[53-55], and these approaches have recently been extended to globular proteins (i.e. β2-microglobulin[50,51,56] and enzymes[49,57,58]).

Residue-specific ncPro mutagenesis. Many of the studies involving the global replacement of proline residues have focused on collagen, a fibrous protein that is abundant in extracellular matrices. Collagen has a consensus sequence $(XaaYaaGly)_n$, where Xaa is L-proline, Yaa is typically (4R)-hydroxy-L-proline (Hyp), and n is the number of repeats[59]. The orientation of the hydroxyl group at position Yaa is critical as the replacement of Hyp with its stereoisomer (4S)-hydroxy-L-proline (Hzp) results in a non-helical structure[60]. It was initially hypothesized that hydrogen bonding was responsible for the helical formation and stability of collagen[61]; however, studies involving replacement at Yaa with (4R)-fluoro-L-proline (Fyp) proved that the preorganization of the triplet repeats was the more important factor[62].

Site-specific ncPro mutagenesis. There are several examples in the literature regarding the reliance of protein structure and function on a single proline residue. One of the first studies to use proline analogs as probes was done on a 5-hydroxytryptamine type 3 ($5\text{-}HT_3$) receptor, containing a transmembrane region structurally and functionally homologous to the nicotinic acetylcholine receptor (nAChR) where a conserved proline (Pro8*) is located[63]. It was determined, through the use of several proline analogs, that the molecular mechanism for ion-gating relied on the *cis-trans* isomerization preferences of Pro8* and its functionality depended on the equilibrium between the *cis* and *trans* conformations of the proline peptide bond. The bulk of this thesis (Chapters 2-4) will describe an analogous

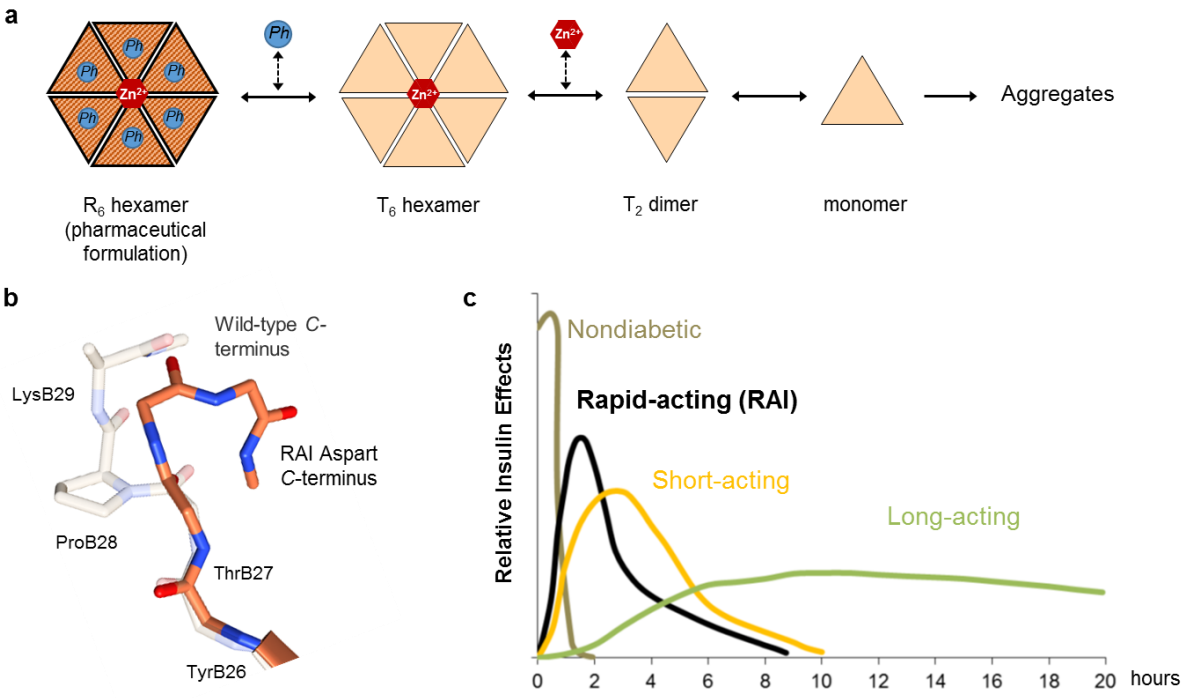
approach to understand the biophysics of the endocrine hormone (and important therapeutic) insulin and its implications for broadening the scope of protein engineering.

Insulin. Blood glucose levels are tightly controlled in mammals through a sensitive regulatory system mediated by insulin, a 51-amino acid endocrine hormone composed of two disulfide-linked polypeptide chains (designated A and B). Upon binding to its receptor, insulin initiates a signaling cascade that accelerates glucose uptake and glycogen production. In diabetic patients, this system malfunctions, and glucose levels must be controlled through subcutaneous injections of insulin[64]. Diabetes is a chronic illness with two types: Type 1, where the patient's β cells are destroyed due to autoimmunity and therefore, cannot produce insulin, and Type 2, where the patient becomes insulin-resistant. Currently, there are millions of people suffering from diabetes worldwide and this number has been projected to rise drastically in the next decade[65]—therefore, there exists a large unmet need that warrants the continued research and development of insulin therapy. This is because currently available prandial insulins are suboptimal in at least two respects. First, the onset of action after injection is delayed by slow dissociation of the insulin hexamer in the subcutaneous space, and second, insulin forms amyloid fibrils upon storage in solution[64]. Insulin can be stabilized with respect to fibrillation by addition of zinc and phenolic preservatives, which drive assembly of the R_6 hexamer (Figure 1.1a)[66-68]. Each subunit in the insulin hexamer adopts one of two conformational states (T or R), depending on the concentration of phenolic ligand[69]. Pharmaceutical formulations are prepared in the more stable R_6 form, whereas the T-state is observed in the absence of phenolic ligands, most commonly in the form of T_2-dimers[70] (Figure 1.1a).

Unfortunately, the R_6 form of insulin is inactive and the lag time for dissociation delays the onset of action[71]. The *C*-terminus of the B-chain is important in mediating dimerization of insulin[72,73], and the flexibility of the B-chain *C*-terminus is believed to contribute to aggregation through formation of amyloid fibrils[74-76].

Figure 1.1 | Oligomeric and *in vivo* behavior of insulin and therapeutic insulins.

***a**, Schematic of hexamer disassembly (adapted from mechanism previously described[67]). Phenolic ligand (Ph), zinc ion (Zn^{2+}), insulin monomer (triangle). Darker shading indicates the R-state of the hexamer. **b**, Structures of the B-chain C-termini of wild-type insulin (ProI, 1ZNJ) and RAI Aspart (AspI, 1EV3). **c**, Relative insulin effects (curves not to scale) of different categories of therapeutic insulin (adapted from figure previously described[66]).*



Rapid-acting insulins (RAIs). Mutation of ProB28 yields RAIs (i.e. RAI Aspart with AspB28, RAI Lispro with LysB28 and ProB29) by disrupting contacts that are critical for dimer formation[67] and consequently leads to hexamer destabilization, which increases the rate of disassembly. Thus, the onset of action in diabetic patients for RAIs requires about half the

time required for regular (short-acting) insulins. There are drawbacks to current RAIs: replacement of Pro through conventional mutagenesis increases the flexibility and perturbs the trajectory of the protein backbone (Figure 1.1b). Proline's dihedral angles are conformationally constrained due to its cyclic structure while all of the other 19 cAAs are acyclic, indicating that no cAA (other than proline) can accurately mimic the proline amide bond. We sought a means to disrupt the dimer interface without releasing the conformational constraints characteristic of proline by using ncAA mutagenesis[52,69,77].

Long-acting (basal) insulins. Normal functioning β cells are constantly secreting insulin at low levels[78]; to mimic this biological process, diabetics inject basal insulins once- or twice-a-day in addition to either (or both) RAIs and regular insulins just prior to mealtime[79]. Basal insulins are made through either conjugation with fatty acids for binding to albumin or other serum proteins in the blood[79], or formulated differently[27] (e.g. insulin-protamine, NPH, is formulated with protamine to reduce solubility). Currently, the longest-acting basal insulin is insulin degludec (marketed as Tresiba), and depending on the patient it is injected either daily[80] or thrice weekly[81].

Unmet needs and future outlooks. Even though insulin is one of the most studied and utilized proteins to date, there is room for improvements to be made. For example, an extremely undesirable side effect of insulin therapy is weight gain[82] leading to interest in developing a 'weight-neutral' insulin therapy. An area of active interest comes from the observation that in non-diabetics, most of the insulin secreted is directed and cleared by the liver (where glycogenesis occurs)[83]. Current insulin therapies[84] demonstrate a systemic uptake instead; the only preferentially hepato-specific insulin to date was PEGylated insulin

lispro[85], made by Eli Lilly through conjugating a large polyethylene glycol (PEG) molecule at LysB29 through amine coupling. However, PEG-lispro was discontinued during Phase III clinical trials due to hepatic toxicity (presumably due to PEG accumulation[86]). Current insulin therapies also greatly rely on patient compliance (monitoring blood glucose levels and injecting insulin prior to meals) due to the lag time for onset of action, which has led to an unmet need for the development of glucose-responsive therapies. Current research for glucose-response therapies are focused on engineering novel insulin formulations to become activated in the presence of high blood glucose and withstand autonomous insulin pump use. Use of pumps with currently available insulins has been plagued by stability issues[87,88]; therefore, there is continued interest in further stabilizing insulins for long-term use in pumps[89,90]. The bulk of the work with glucose-responsive insulin formulations has focused on delivery vehicles[91-93] that have not been validated clinically.

Demonstrating how the therapeutic and physical properties of insulin can be modulated with ncAAs and expanding on methods to utilize ncAAs will give scientists another powerful new tool, one that fuses the concepts of medicinal chemistry and protein design, for the design of antibody-drug conjugates, bispecific antibodies, and other novel protein therapeutics.

References

1. Drummond, D.A. & Wilke, C.O. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet* **10**(10): 715-724 (2009).
2. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* **353**(2005).
3. Steward, R.E., MacArthur, M.W., Laskowski, R.A. & Thornton, J.M. Molecular basis of inherited diseases: a structural perspective. *Trends in Genetics* **19**(9): 505-513 (2003).

4. Chiti, F. & Dobson, C.M. Protein misfolding, functional amyloid, and human disease. *Annu Rev. Biochem* **75**(1): 333-366 (2006).
5. Zimmet, P., Alberti, K.G.M.M. & Shaw, J. Global and societal implications of the diabetes epidemic. *Nature* **414**(6865): 782-787 (2001).
6. Leader, B., Baca, Q.J. & Golan, D.E. Protein therapeutics: a summary and pharmacological classification. *Nat. Rev. Drug Discov* **7**(1): 21-39 (2008).
7. Rosenfeld, L. Insulin: discovery and controversy. *Clinical Chemistry* **48**(12): 2270 (2002).
8. Pavlou, A.K. & Reichert, J.M. Recombinant protein therapeutics[mdash]success rates, market trends and values to 2010. *Nat Biotech* **22**(12): 1513-1519 (2004).
9. Turner, K.B., Alves, N.J., Medintz, I.L. & Walper, S.A. Improving the targeting of therapeutics with single-domain antibodies. *Expert Opinion on Drug Delivery* **13**(4): 561-570 (2016).
10. Rodgers, K.R. & Chou, R.C. Therapeutic monoclonal antibodies and derivatives: Historical perspectives and future directions. *Biotechnol Adv.* **34**(6): 1149-1158 (2016).
11. Harris, J.M. & Chess, R.B. Effect of pegylation on pharmaceuticals. *Nat. Rev. Drug Discov* **2**(3): 214-21 (2003).
12. Frokjaer, S. & Otzen, D.E. Protein drug stability: a formulation challenge. *Nat. Rev. Drug Discov* **4**(4): 298-306 (2005).
13. Qi, Y. & Chilkoti, A. Protein–polymer conjugation — moving beyond PEGylation. *Curr Opin Chem Biol* **28**: 181-193 (2015).
14. Zelikin, A.N., Ehrhardt, C. & Healy, A.M. Materials and methods for delivery of biological drugs. *Nat Chem* **8**(11): 997-1007 (2016).
15. Renaud, J.-P. et al. Biophysics in drug discovery: impact, challenges and opportunities. *Nat. Rev. Drug Discov* **15**: 679-698 (2016).
16. Schmidt, S., Gonzalez, D. & Derendorf, H. Significance of protein binding in pharmacokinetics and pharmacodynamics. *J Pharm Sci* **99**(3): 1107-1122 (2010).
17. Li, Y. et al. Variant fatty acid-like molecules conjugation, novel approaches for extending the stability of therapeutic peptides. *Scientific Reports* **5**: 18039 (2015).
18. Shire, S.J., Shahrokh, Z. & Liu, J. Challenges in the development of high protein concentration formulations. *J Pharm Sci* **93**(6): 1390-1402 (2004).
19. Cleland, J.L. & Langer, R. Formulation and delivery of proteins and peptides. in *Formulation and Delivery of Proteins and Peptides*, Vol. 567 1-19 (American Chemical Society, 1994).
20. Thirumalai, D. & Reddy, G. Protein thermodynamics: Are native proteins metastable? *Nat Chem* **3**(12): 910-911 (2011).
21. Gazit, E. The “correctly folded” state of proteins: is it a metastable state? *Angew Chem Int Ed Engl* **41**(2): 257-259 (2002).
22. Mitragotri, S., Burke, P.A. & Langer, R. Overcoming the challenges in administering biopharmaceuticals: formulation and delivery strategies. *Nat. Rev. Drug Discov* **13**(9): 655-672 (2014).
23. Manning, M.C., Chou, D.K., Murphy, B.M., Payne, R.W. & Katayama, D.S. Stability of protein pharmaceuticals: An update. *Pharm. Res.* **27**(4): 544-575 (2010).

24. Librizzi, F. & Rischel, C. The kinetic behavior of insulin fibrillation is determined by heterogeneous nucleation pathways. *Protein Sci.* **14**(12): 3129-3134 (2005).
25. Shire, S.J. Formulation and manufacturability of biologics. *Current Opinion in Biotechnology* **20**(6): 708-714 (2009).
26. Hassiepen, U., Federwisch, M., Mulders, T. & Wollmer, A. The lifetime of insulin hexamers. *Biophys. J* **77**(3): 1638-54 (1999).
27. Gualandi-Signorini, A.M. & Giorgi, G. Insulin formulations -- a review. *Eur Rev Med Pharmacol Sci* **5**(3): 73-83 (2001).
28. Peyrot, M., Barnett, A.H., Meneghini, L.F. & Schumm-Draeger, P.M. Insulin adherence behaviours and barriers in the multinational Global Attitudes of Patients and Physicians in Insulin Therapy study. *Diabetic Medicine* **29**(5): 682-689 (2012).
29. 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* **14**(6): 774-80 (2010).
30. 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 Engl* **39**(12): 2148-2152 (2000).
31. Dougherty, D.A. Unnatural amino acids as probes of protein structure and function. *Curr Opin Chem Biol* **4**(6): 645-652 (2000).
32. Davis, L. & Chin, J.W. Designer proteins: applications of genetic code expansion in cell biology. *Nat Rev Mol Cell Biol* **13**(3): 168-82 (2012).
33. Johnson, D.B. et al. RF1 knockout allows ribosomal incorporation of unnatural amino acids at multiple sites. *Nat Chem Biol* **7**(11): 779-86 (2011).
34. Amiram, M. et al. Evolution of translation machinery in recoded bacteria enables multi-site incorporation of nonstandard amino acids. *Nat Biotech* **33**(12): 1272-1279 (2015).
35. Lajoie, M.J. et al. Genomically recoded organisms expand biological functions. *Science* **342**(6156): 357 (2013).
36. Wals, K. & Ovaa, H. Unnatural amino acid incorporation in E. coli: current and future applications in the design of therapeutic proteins. *Frontiers in Chemistry* **2**: 15 (2014).
37. Carlson, E.D., Gan, R., Hodgman, C.E. & Jewett, M.C. Cell-free protein synthesis: Applications come of age. *Biotechnol Adv.* **30**(5): 1185-1194 (2012).
38. Spicer, C.D. & Davis, B.G. Selective chemical protein modification. *Nat Commun* **5**: 4740 (2014).
39. Link, A.J. & Tirrell, D.A. Cell surface labeling of *Escherichia coli* via copper(I)-catalyzed [3+2] cycloaddition. *JACS* **125**(37): 11164-11165 (2003).
40. 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. U. S. A.* **103**(25): 9482-7 (2006).
41. Kiick, K.L., Saxon, E., Tirrell, D.A. & Bertozzi, C.R. Incorporation of azides into recombinant proteins for chemoselective modification by the Staudinger ligation. *Proc. Natl. Acad. Sci. U. S. A.* **99**(1): 19-24 (2002).

42. Chahrour, O., Cobice, D. & Malone, J. Stable isotope labelling methods in mass spectrometry-based quantitative proteomics. *Journal of Pharmaceutical and Biomedical Analysis* **113**: 2-20 (2015).
43. Bakalarski, C.E. & Kirkpatrick, D.S. A Biologist's Field Guide to Multiplexed Quantitative Proteomics. *Molecular & Cellular Proteomics* **15**(5): 1489-1497 (2016).
44. Lu, Y.Y. et al. Pro-metastatic GPCR CD97 is a Direct Target of Tumor Suppressor microRNA-126. *ACS Chem Biol* **9**(2): 334-338 (2014).
45. Yuet, K.P. et al. Cell-specific proteomic analysis in Caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.* **112**(9): 2705-2710 (2015).
46. Bagert, J.D. et al. Time-resolved proteomic analysis of quorum sensing in Vibrio harveyi. *Chemical Science* **7**(3): 1797-1806 (2016).
47. Tang, Y. et al. Stabilization of Coiled-Coil Peptide Domains by Introduction of Trifluoroleucine. *Biochemistry* **40**(9): 2790-2796 (2001).
48. Milner-White, E.J., Bell, L.H. & Maccallum, P.H. Pyrrolidine ring puckering in cis and trans-proline residues in proteins and polypeptides. *J Mol Biol* **228**(3): 725-734 (1992).
49. 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* **14**(9): 1053-1057 (2013).
50. Torbeev, V.Y. & Hilvert, D. Both the cis-trans equilibrium and isomerization dynamics of a single proline amide modulate β2-microglobulin amyloid assembly. *Proc. Natl. Acad. Sci. U. S. A.* **110**(50): 20051-20056 (2013).
51. Torbeev, V., Ebert, M.-O., Dolenc, J. & Hilvert, D. Substitution of proline32 by α-methylproline preorganizes β2-microglobulin for oligomerization but not for aggregation into amyloids. *JACS* (2015).
52. Pandey, A.K., Naduthambi, D., Thomas, K.M. & Zondlo, N.J. Proline editing: a general and practical approach to the synthesis of functionally and structurally diverse peptides. Analysis of steric versus stereoelectronic effects of 4-substituted prolines on conformation within peptides. *JACS* **135**(11): 4333-4363 (2013).
53. Hodges, J.A. & Raines, R.T. Stereoelectronic effects on collagen stability: The dichotomy of 4-fluoroproline diastereomers. *JACS* **125**(31): 9262-9263 (2003).
54. Shoulders, M.D., Kamer, K.J. & Raines, R.T. Origin of the stability conferred upon collagen by fluorination. *Bioorg. Med. Chem. Lett.* **19**(14): 3859-3862 (2009).
55. Shoulders, M.D. & Raines, R.T. Interstrand dipole-dipole interactions can stabilize the collagen triple helix. *J. Biol. Chem.* **286**(26): 22905-22912 (2011).
56. 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* **6**(5): e19425 (2011).
57. Roderer, D., Glockshuber, R. & Rubini, M. Acceleration of the rate-limiting step of Thioredoxin folding by replacement of its conserved cis-proline with (4S)-fluoroproline. *ChemBioChem* **16**(15): 2162-2166 (2015).
58. Holzberger, B., Rubini, M., Möller, H.M. & Marx, A. A highly active DNA polymerase with a fluorous core. *Angew Chem Int Ed Engl* **49**(7): 1324-1327 (2010).

59. Bretscher, L.E., Jenkins, C.L., Taylor, K.M., DeRider, M.L. & Raines, R.T. Conformational stability of collagen relies on a stereoelectronic effect. *JACS* **123**(4): 777-778 (2001).
60. Shoulders, M.D., Kotch, F.W., Choudhary, A., Guzei, I.A. & Raines, R.T. The aberrance of the 4*S* diastereomer of 4-hydroxyproline. *JACS* **132**(31): 10857-10865 (2010).
61. Shoulders, M.D. & Raines, R.T. Collagen structure and stability. *Annu Rev. Biochem* **78**: 929-58 (2009).
62. Holmgren, S.K., Bretscher, L.E., Taylor, K.M. & Raines, R.T. A hyperstable collagen mimic. *Chemistry & Biology* **6**(2): 63-70 (1999).
63. Lummis, S.C.R. et al. Cis-trans isomerization at a proline opens the pore of a neurotransmitter-gated ion channel. *Nature* **438**(7065): 248-252 (2005).
64. Zaykov, A.N., Mayer, J.P. & DiMarchi, R.D. Pursuit of a perfect insulin. *Nat. Rev. Drug Discov* **15**: 425-439 (2016).
65. Zimmet, P.Z., Magliano, D.J., Herman, W.H. & Shaw, J.E. Diabetes: a 21st century challenge. *The Lancet Diabetes & Endocrinology* **2**(1): 56-64 (2014).
66. Freeman, J.S. Insulin analog therapy: improving the match with physiologic insulin secretion. *J. Am. Osteopath. Assoc.* **109**(1): 26-36 (2009).
67. Birnbaum, D.T., Kilcomons, M.A., DeFelippis, M.R. & Beals, J.M. Assembly and dissociation of human insulin and LysB28ProB29-insulin hexamers: a comparison study. *Pharm. Res.* **14**(1): 25-36 (1997).
68. Carpenter, M.C. & Wilcox, D.E. Thermodynamics of formation of the insulin hexamer: metal-stabilized proton-coupled assembly of quaternary structure. *Biochemistry* **53**(8): 1296-301 (2014).
69. Pandyarajan, V. & Weiss, M.A. Design of non-standard insulin analogs for the treatment of diabetes mellitus. *Curr. Diab. Rep.* **12**(6): 697-704 (2012).
70. Palmieri, L.C., Favero-Retto, M.P., Lourenco, D. & Lima, L.M. A T3R3 hexamer of the human insulin variant B28Asp. *Biophys. Chem.* **173-174**: 1-7 (2013).
71. Bakaysa, D.L. et al. Physicochemical basis for the rapid time-action of LysB28ProB29-insulin: dissociation of a protein-ligand complex. *Protein Sci.* **5**(12): 2521-2531 (1996).
72. Ciszak, E. et al. Role of C-terminal B-chain residues in insulin assembly: the structure of hexameric LysB28ProB29-human insulin. *Structure* **3**(6): 615-22 (1995).
73. Menting, J.G. et al. How insulin engages its primary binding site on the insulin receptor. *Nature* **493**(7431): 241-5 (2013).
74. Huang, K., Maiti, N.C., Phillips, N.B., Carey, P.R. & Weiss, M.A. Structure-specific effects of protein topology on cross-β assembly: studies of insulin fibrillation. *Biochemistry* **45**(34): 10278-10293 (2006).
75. Menting, J.G. et al. Protective hinge in insulin opens to enable its receptor engagement. *Proc. Natl. Acad. Sci. U. S. A.* **111**(33): E3395-404 (2014).
76. Ivanova, M.I., Sievers, S.A., Sawaya, M.R., Wall, J.S. & Eisenberg, D. Molecular basis for insulin fibril assembly. *Proc. Natl. Acad. Sci. U. S. A.* **106**(45): 18990-18995 (2009).
77. Liu, C.C. & Schultz, P.G. Adding new chemistries to the genetic code. *Annu Rev. Biochem* **79**: 413-44 (2010).

78. Wilcox, G. Insulin and insulin resistance. *Clinical Biochemist Reviews* **26**(2): 19-39 (2005).
79. Jonassen, I. et al. Design of the novel protraction mechanism of insulin degludec, an ultra-long-acting basal insulin. *Pharm. Res.* **29**(8): 2104-2114 (2012).
80. Wang, F., Surh, J. & Kaur, M. Insulin degludec as an ultralong-acting basal insulin once a day: a systematic review. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* **5**: 191-204 (2012).
81. Zinman, B. et al. Efficacy and safety of insulin degludec three times a week versus insulin glargine once a day in insulin-naive patients with type 2 diabetes: results of two phase 3, 26 week, randomised, open-label, treat-to-target, non-inferiority trials. *The Lancet Diabetes & Endocrinology* **1**(2): 123-131 (2013).
82. Caparrotta, T.M. & Evans, M. PEGylated insulin lispro, (LY2605541)—a new basal insulin analogue. *Diabetes, Obesity and Metabolism* **16**(5): 388-395 (2014).
83. Edgerton, D.S. et al. Insulin's direct effects on the liver dominate the control of hepatic glucose production. *The Journal of Clinical Investigation* **116**(2): 521-527 (2006).
84. Meier, J.J., Veldhuis, J.D. & Butler, P.C. Pulsatile insulin secretion dictates systemic insulin delivery by regulating hepatic insulin extraction in humans. *Diabetes* **54**(6): 1649-1656 (2005).
85. Madsbad, S. LY2605541—A preferential hepato-specific insulin analogue. *Diabetes* **63**(2): 390-392 (2014).
86. Muñoz-Garach, A., Molina-Vega, M. & Tinahones, F.J. How Can a Good Idea Fail? Basal Insulin Peglispro [LY2605541] for the Treatment of Type 2 Diabetes. *Diabetes Therapy*: 1-14 (2016).
87. Teska, B.M., Alarcón, J., Pettis, R.J., Randolph, T.W. & Carpenter, J.F. Effects of phenol and *m*-cresol depletion on insulin analog stability at physiological temperature. *J Pharm Sci* **103**(8): 2255-2267 (2014).
88. Bode, B. Comparison of pharmacokinetic properties, physicochemical stability, and pump compatibility of 3 rapid-acting insulin analogues-aspart, lispro, and glulisine. *Endocrine Practice* **17**(2): 271-280 (2010).
89. Vinther, T.N. et al. Additional disulfide bonds in insulin: Prediction, recombinant expression, receptor binding affinity, and stability. *Protein Sci.* **24**(5): 779-88 (2015).
90. Karas, J.A. et al. Total Chemical Synthesis of an Intra-A-Chain Cystathionine Human Insulin Analogue with Enhanced Thermal Stability. *Angew Chem Int Ed Engl* **55**(47): 14743-14747 (2016).
91. Tai, W. et al. Bio-Inspired synthetic nanovesicles for glucose-responsive release of insulin. *Biomacromolecules* **15**(10): 3495-3502 (2014).
92. Yu, J. et al. Microneedle-array patches loaded with hypoxia-sensitive vesicles provide fast glucose-responsive insulin delivery. *Proc. Natl. Acad. Sci. U. S. A.* **112**(27): 8260-8265 (2015).
93. Gu, Z. et al. Glucose-Responsive Microgels Integrated with Enzyme Nanocapsules for Closed-Loop Insulin Delivery. *ACS Nano* **7**(8): 6758-6766 (2013).

Chapter 2 – Stereo-specific hydroxylation of proline at position B28 of insulin

Abstract

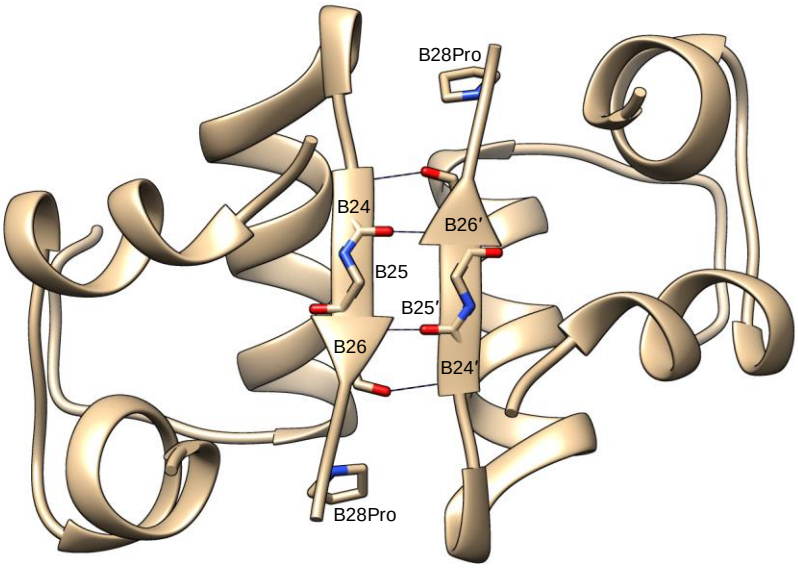
Rapid-acting insulins (RAIs) have been produced by mutation of the proline residue at position 28[1,2] of the insulin B-chain (ProB28); mutation at this position disrupts the inter-subunit interface in the hexamer and enhances the rate of disassembly, but it does not delay fibrillation. Here we show that replacement of ProB28 by (*4S*)-hydroxyproline (Hzp) yields an active form of insulin that dissociates more rapidly, and fibrillates more slowly, than the wild-type protein. In contrast to the behavior of known RAIs[3], the rapid dissociation of Hzp-insulin (HzpI) is not accompanied by substantial destabilization of the insulin dimer; instead, a novel hydrogen bond interaction was found in the crystal structure of HzpI. The effects of hydroxylation are stereospecific; replacement of ProB28 by (*4R*)-hydroxyproline (Hyp) causes little change in the rates of fibrillation and hexamer disassembly but disrupts dimerization.

Introduction

Insulin embodies a conundrum that often occurs with proteins: tradeoffs between stability and activity. Evolution has adapted proteins to behave a certain way to fulfill specific functions. For example, insulin is secreted when it is needed (to lower blood glucose levels) and is rapidly removed when it is not (when blood glucose reaches homeostasis)[4]. The insulin-producing and -secreting β cells in the islets of Langerhans, which are located in the pancreas, are able to lower high blood glucose concentrations to normal levels in a manner of minutes[5,6]. In order to elicit such a quick reaction, at any given time, hundreds of thousands of insulin molecules are stored in secretory granules[5], but to avoid hypoglycemia, the unstable insulin monomer is also rapidly removed from circulation (by insulinase)[7] once it is used.

Figure 2.1 | Structure of dimeric insulin.

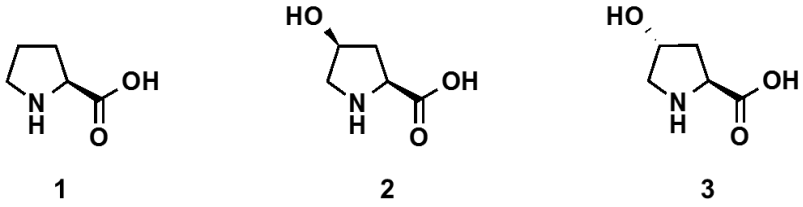
Crystal structure of WT insulin (PDB: 3T2A) with residues involved in forming the dimer interface labelled. Lines represent hydrogen bonds (analyzed using UCSF Chimera software) between Residues B24-B26 of each insulin monomer.



The oligomeric behavior of insulin is due to the inter-subunit interactions[8] (Figure 2.1); for example, a micromolar solution of insulin, in the absence of any metal ions or phenolic ligands, is primarily dimeric[9]. However, at millimolar concentrations, insulin exists in mostly hexameric form[10]. Decades of insulin studies (i.e. mutagenesis at every residue) have determined that the oligomeric behavior of insulin is sensitive to perturbations at the C-terminus of the B chain (particularly residues B26-B30)[2,11,12].

Figure 2.2 | Proline analogs for study of hydroxylation

Compound **1***: L-proline* (Pro)*; Compound* **2***: (2S,4S)-4-hydroxy-L-proline* (Hzp)*; Compound* **3***: (2S,4R)-4-hydroxy-L-proline* (Hyp)



Hydroxy-prolines. Previous work has shown that canonical mutations of ProB28 to acyclic, charged amino acids (i.e. aspartic acid, lysine) can modulate the dimerization of insulin[1,13]. However, acyclic amino acids can access more conformational space than proline[14], which results in additional flexibility of the polypeptide backbone. Therefore, proline analogs (i.e. hydroxyprolines) were chosen to take advantage of a polar functional group with the added benefit of being able to maintain proline backbone trajectory[15] because of the restricted dihedral angles. In addition to their polarity and subsequent capacity for hydrogen-bonding, hydroxylation at the 4th position alters the *endo*/*exo* preference of the pyrrolidine ring and the *cis*/*trans* equilibrium of the amide backbone[16-18]. Calculations of *cis/trans* equilibrium

constants for Hzp and Hyp using model peptides show Hzp promotes a *cis*-amide, and Hyp, a *trans*-amide peptide bond[19].

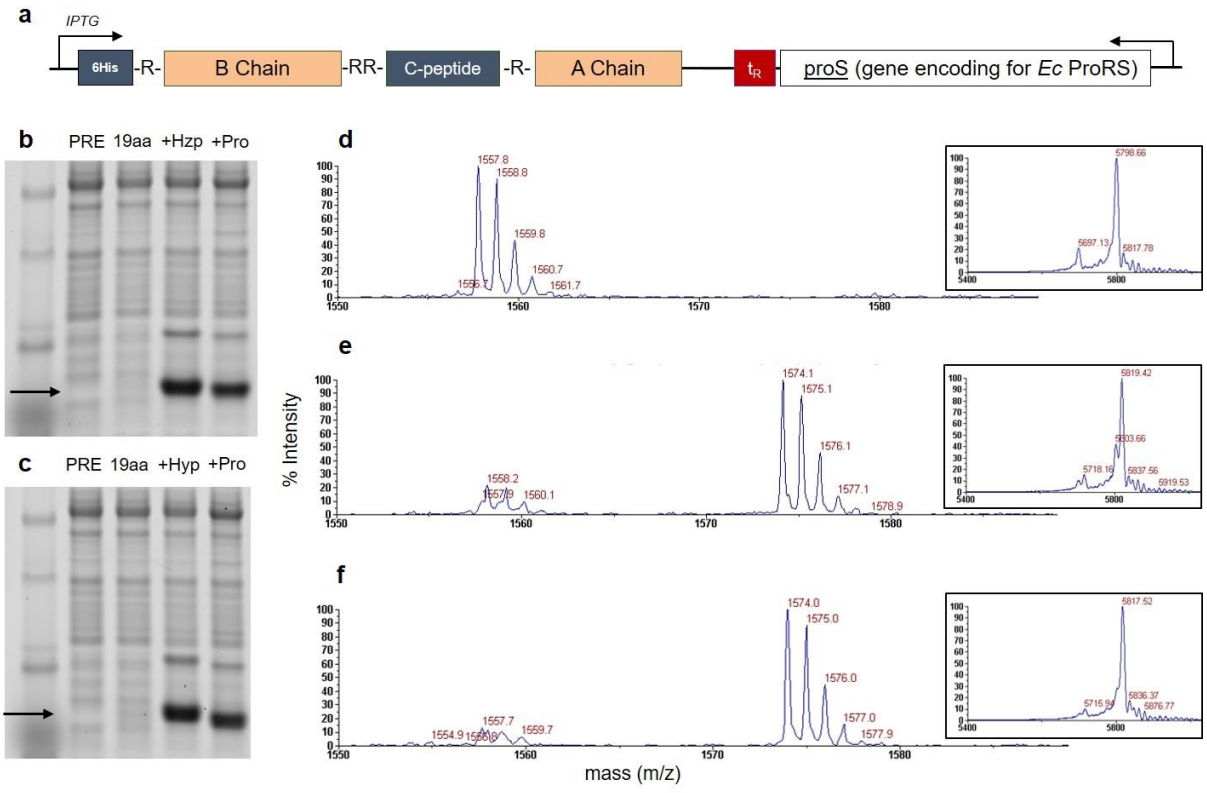
Results and Discussion

Recombinant expression of hydroxyinsulins. Non-diabetics produce insulin as a cleavage product from preproinsulin[20] consisting of a signal peptide, followed by a B chain and A chain that are connected by a C-peptide. This process has been adapted for the use of *E. coli* as an expression host starting with the expression of proinsulin (PI, preproinsulin without the signal peptide) followed by *in vitro* refolding and cleavage to mature insulin[20,21]. It has been previously reported[22] that Hzp and Hyp can be incorporated into newly synthesized proteins in *E. coli* (*Ec)* with an overexpressed prolyl-tRNA synthetase (ProRS) under osmotic stress conditions. It is known that proline is an effective osmoprotectant and proline transporters are upregulated in response to osmotic stress (i.e. NaCl at concentrations >300 mM)[23]. PI was placed under the inducible *lac* promoter on expression vector pQE80L with the gene encoding for *Ec* ProRS under its endogenous proS promoter located downstream between two transcriptional termination sites (pQE80PI-proS; Figure 2.3a). Expression was done using proline-auxotrophic *E. coli* (strain CAG18515), and grown to mid-exponential phase before it was subjected to a medium shift to minimal medium with an extended depletion step to aid in the removal of intracellular proline and help enhance ncPro incorporation by selective pressure. After depletion, Hzp or Hyp, in the presence of 300 mM NaCl was added to allow for Hzp or Hyp uptake before cultures were induced with isopropyl-β-D-thiogalactopyranoside (IPTG) for 2 h. Expression of PI was confirmed via polyacrylamide gel electrophoresis (Figure 2.3bc). Incorporation levels were

assessed using matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS; Figure 2.3 d-f) and found to be approximately 90% by comparing areas of the peaks corresponding to the hydroxylated peptide at m/z = 1574 Da and the wild-type peptide at m/z = 1558 Da.

Figure 2.3 | Insulin expression and incorporation of hydroxyprolines.

a, *Plasmid pQE80PI-proS containing gene construct proinsulin under inducible lac promoter and an overexpressed Ec ProRS for bacterial expression.* ***b, c,*** *SDS-PAGE of cell lysates with lanes labeled for pre-induction (PRE) and post-induction in minimal media supplemented with either nothing (19aa), Hzp (**b**) or Hyp (**c**), or Pro at 0.5 mM.* ***c-e:*** *MALDI-MS traces of isolated proinsulin peptide fragment ^{46}RGFFYT**P**KTRRE^{57} obtained by gluC digestion. Peptide fragment masses corresponds to either wild type mass 1558 Da (**d**) or shifted mass 1574 Da if Hzp (**e**) or Hyp (**f**) is incorporated. Inset is whole protein MALDI-MS. Note the presence of desB30 insulin is <10%. All MALDI-MS spectra contain ion counts >10^3*

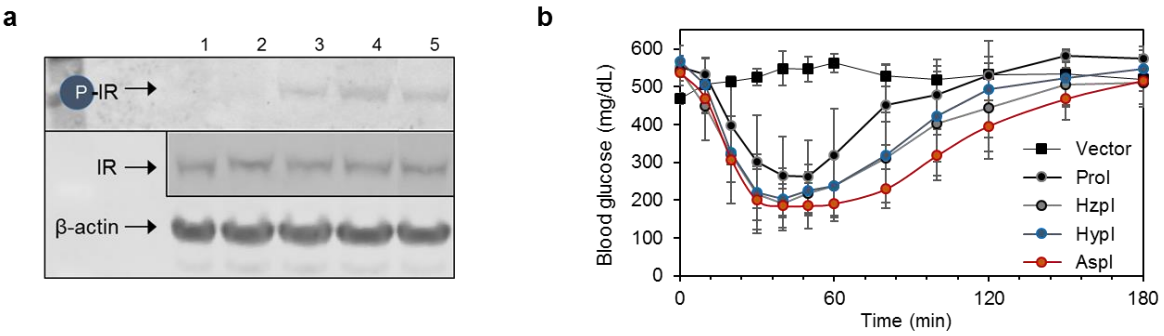


Refolding and insulin maturation. Inclusion bodies containing PI were extracted and resolubilized in denaturing buffer and purified using Ni-NTA affinity chromatography.

Elution fractions containing PI were combined, pH-adjusted, and subjected to oxidative sulfitolysis, and refolded. Typical soluble protein yields were ~30 mg/L for PI expressed in minimal media. The refolded protein was isolated by shifting to the solution to acidic conditions and then dialyzed, lyophilized and re-suspended in water for proteolytic cleavage (with trypsin and carboxypeptidase B) and liquid chromatography (HPLC) purification. Post-HPLC fractions containing mature insulin were lyophilized, re-suspended in phosphate buffer prior to verification by whole protein MALDI-MS (Figure 2.3c-e, inset). Recovery of insulin from soluble PI was estimated to be 25-50%. Insulin fractions were stored at -80°C and thawed on ice prior to use.

Figure 2.4 | Hydroxyinsulins retain biological activity.

a, *Immunoblot of phosphorylated insulin receptor using HEK293 cells treated with insulin (200 nM in PBS, pH 7.4) or vehicle. Whole cell lysates were then run on an SDS-PAGE gel and transferred to nitrocellulose membrane to detect insulin receptor (IR) and IR phosphorylation. β-actin immunoblot shown as loading control. Lane 1: Vehicle (PBS); Lane 2: 10% ProI serving as a second negative control due to presence of 10% wt in HzpI and HypI preparations; Lane 3: HzpI; Lane 4: HypI; Lane 5: ProI.* ***b,*** *Reduction of blood glucose following subcutaneous injection of 35 μg/kg insulins into streptozocin-induced diabetic mice. Glucose levels were measured post-injection via tail vein sampling. ProI, AspI, HzpI, HypI or vector were formulated as described[24]. Error bars denote one standard deviation (n = 3).*



Recombinant HzpI and HypI are biologically active. We verified that recombinant insulins can bind to the insulin receptor (IR) and initiate its signaling cascade (*via* phosphorylated IR)

in an *in vitro* cell culture assay[25] (Figure 2.4a). Recombinant insulins were also able to lower blood glucose levels in streptozotocin (STZ)-induced diabetic mice (Figure 2.4b) *in vivo*; the presence of 10% ProI in HypI and HzpI preparations (Figure 2.3de, Table 2.1) is not enough to initiate blood glucose lowering action (data shown in Chapter 3, Figure 3.4a). This was expected per literature as residues B26-30 are not involved in IR binding[26]. Recombinant insulin preparations for *in vivo* injections were verified to be endotoxin-free and devoid of common microorganisms (*E.* coli and yeast) by sterility tests.

Table 2.1 | Biophysical characteristics of hydroxyinsulin.

*Errors are given as one standard deviation (n ≥ 4). *Quantified by MALDI-MS on proinsulin peptide obtained by gluC digestion: [46]RGFFYT**P**KTRRE[57]. **Samples contain approximately 10% ProI*

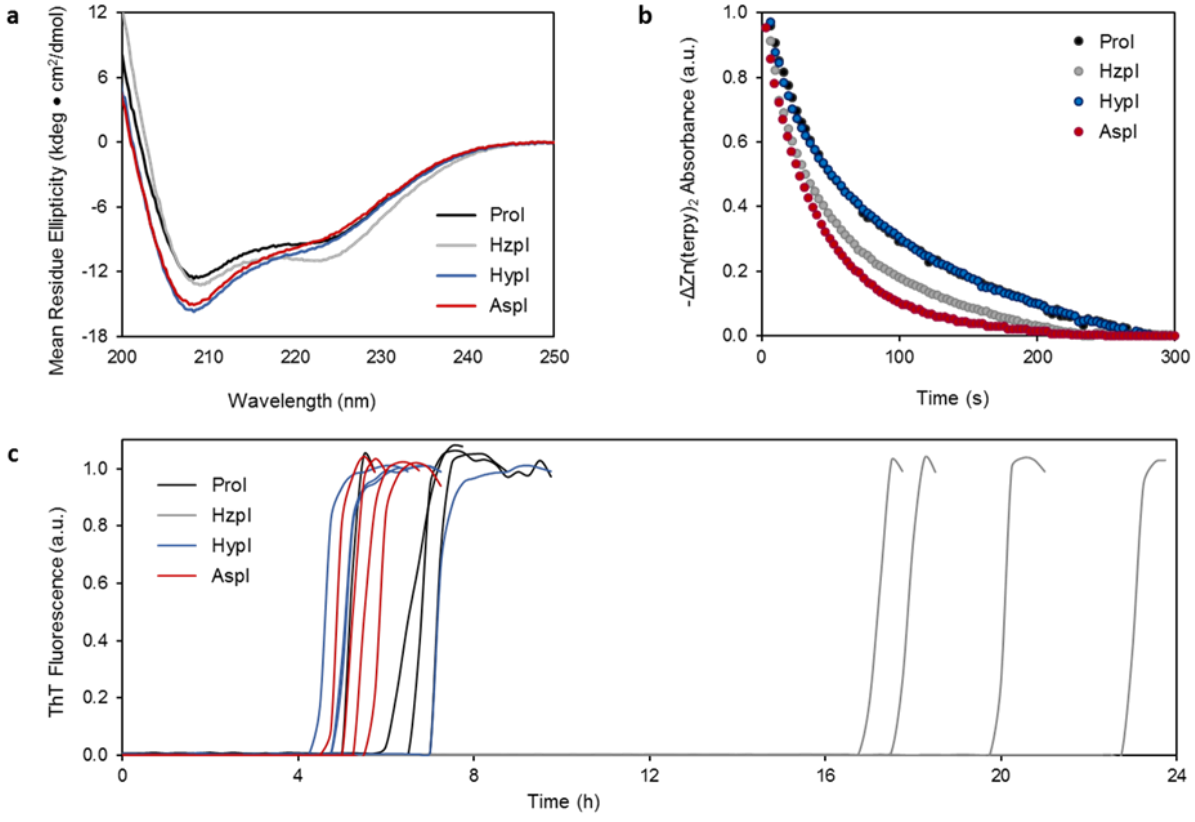
B28 Amino Acid	Incorporation Level* (%)	Insulin Variant	Hexamer $\tau_{1/2}$ (s)	Fibrillation Lag Time (h)	K_D dimer (μM)
L-Proline	--	ProI	90.4 ± 4.2	5.1 ± 1.5	9 μM[27]
Hzp	91.3 ± 1.8	HzpI**	53.6 ± 3.7	19.6 ± 2.6	~25 μM
Hyp	88.2 ± 1.2	HypI**	87.0 ± 10	5.5 ± 1.2	>200 μM
L-Aspartic acid	--	AspI	42.7 ± 4.3	5.3 ± 1.0	>500 μM[13]

Incorporation of hydroxyprolines alters insulin dimerization. In the absence of Zn^{2+} and phenolic preservatives, insulins dimerize with a dissociation constant (K_D) of approximately 10 μM. In contrast, K_D for RAIs is typically >500 μM, and it is believed that destabilization of the dimer interface causes the accelerated onset of insulin action after subcutaneous injection[13,27,28]. Monomeric forms of insulin give rise to characteristic circular dichroism (CD) spectra with distinct minima at 208 and 222 nm (e.g., AspI; Figure 2.5a). Dimerization causes a loss of negative ellipticity at 208 nm (e.g., ProI; Figure 2.5a). At 60 μM, HypI appears to be monomeric (with a CD spectrum nearly identical to that of AspI; Figure 2.5a) while the spectrum of HzpI suggests a dimeric insulin (Figure 2.5a). Sedimentation velocity

(SV) and sedimentation equilibrium (SE) experiments were consistent with the results of the CD analysis (data not shown). SE data were fitted to a model of monomer-dimer-hexamer self-association (SEDPHAT)[29,30], and yielded monomer-dimer dissociation constants (K_D) of >200 µM and 25 µM for HypI and HzpI, respectively.

Figure 2.5 | Hydroxylation at ProB28 modulates insulin dimerization, dissociation kinetics, and stability.

a, Far UV CD spectra collected on 60 µM insulins in 10 mM phosphate buffer, pH 8.0 at 25°C. b, Insulin hexamer dissociation following sequestration of Zn^{2+} by terpyridine. Zn^{2+}-(terpy) signal was monitored at 334 nm and fitted to a mono-exponential decay. HzpI and HypI contain 10% ProI. The curves for HypI and ProI are indistinguishable in this plot. c, Representative fibrillation curves for 60 µM insulins (37°C, 960 RPM; n=4). Insulin fibrils were detected by the rise in Thioflavin T (ThT) fluorescence that occurs upon binding to fibrillar aggregates.

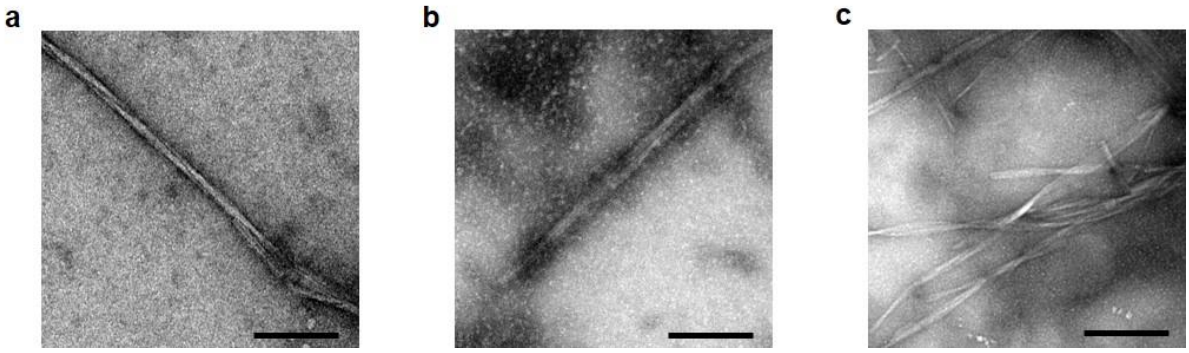


Hydroxylated insulins behave differently from wild-type insulin. Previous studies of RAIs have shown that destabilization of the dimer interface correlates with accelerated

dissociation of the hexamer and rapid onset of insulin action[31,32]. Triggered dissociation of Zn^{2+}-hexamers by addition of the chelating agent terpyridine[33] revealed nearly identical rates of dissociation for HypI and ProI, ($\tau_{1/2}$ = 87.0 ± 10 s and 90.4 ± 4.2 s, respectively; Figure 2.5b) while HzpI exhibited kinetics similar to those of AspI ($\tau_{1/2}$ = 53.6 ± 3.7 s and 42.7 ± 4.3 s, respectively; Figure 2.5b). Replacement by Hyp destabilizes the dimer but has essentially no effect on hexamer dissociation, while introduction of Hzp does not affect dimer stability but enhances rate of hexamer disassembly. Our results indicate that dimerization of the insulin monomer and dissociation kinetics of the insulin hexamer can be decoupled, and thus we speculate that there are other factors, besides destabilization of the dimer, that govern $\tau_{1/2}$.

Figure 2.6 | Micrographs of insulin fibrils.

*Transmission electron micrographs of (**a**) ProI (**b**) HzpI, and (**c**) HypI fibrils. Scale bar 100 nm.*

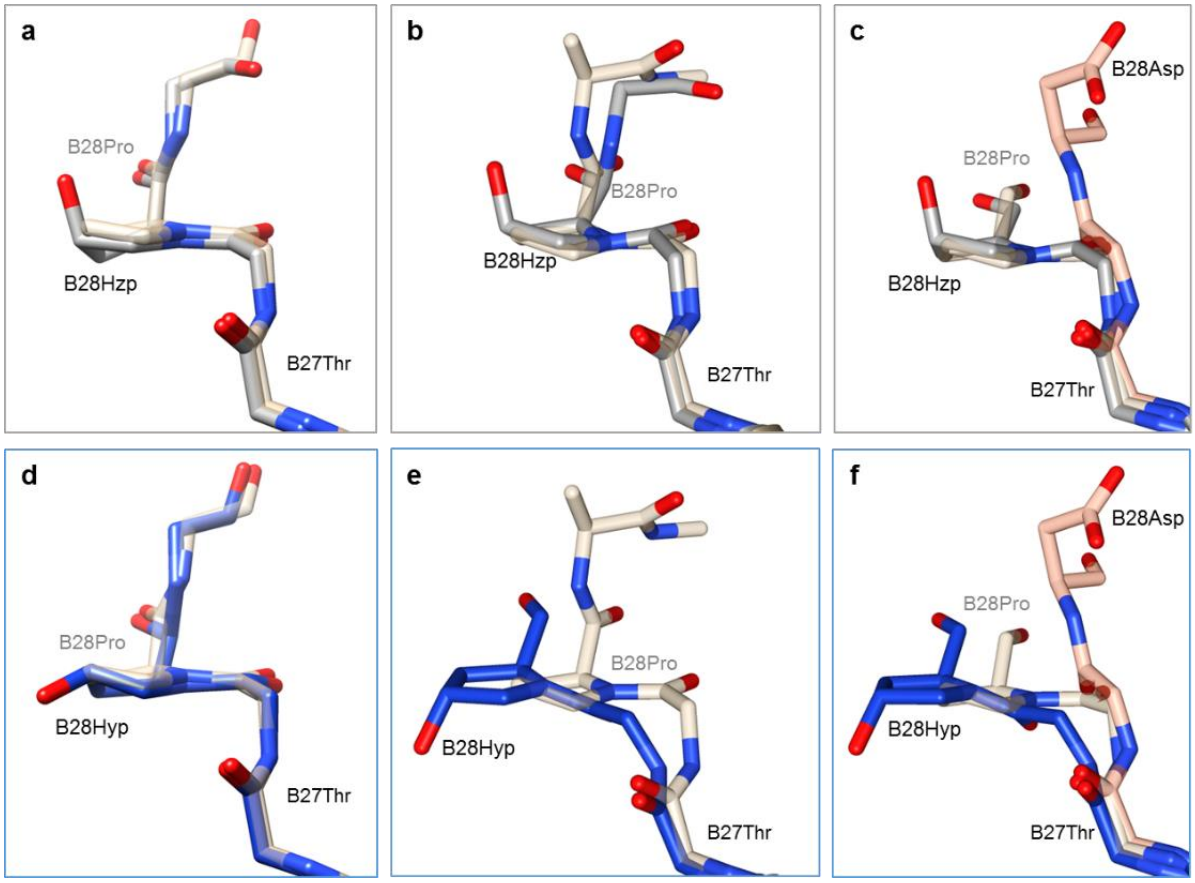


Insulin therapy is associated with amyloidogenesis[34]. While there is a lot known about insulin fibrillation[11,35-46], and some studies have used ncAAs[24,31,47], we cannot predict what effects hydroxylation at position B28 will have—thus, each of the insulin variants was subjected to fibrillation lag time analysis (Figure 2.5c)[48]. We found similar times to onset of fibrillation for HypI, ProI, and AspI; in contrast, HzpI is markedly more resistant to

aggregation. Transmission electron microscopy (TEM) imaging of insulin fibrils found similar structural features (Figure 2.6) between ProI, HzpI, and HypI, indicating that hydroxylation at B28 does not influence the final aggregation state.

Figure 2.7 | Alignment at position B28.

a, d, *Alignment of T_2 ProI (tan, PDB:3T2A), and T_2-HzpI (grey) or T_2-HypI (blue) centered on position B28.* ***b, e,*** *Alignment of R_6-ProI (tan) and R_6-HzpI (grey) or R_6-HypI (blue) highlighting the overlap of the backbone at the C-terminus.* ***c, f,*** *Alignment of R_6 insulins (ProI, and HzpI or HypI), and AspI (orange, PDB: 1ZEG) centered on position B28 highlighting the similarity of the polypeptide backbone for HzpI and HypI, as opposed to AspI, compared to ProI.*



Crystal structure of HzpI reveals a novel hydrogen bond in the dimer interface. To elucidate the molecular origins of the dissociation and fibrillation behavior of HypI and HzpI, we examined crystal structures of both T- and R-states. Hydroxylation at ProB28 does not cause substantial perturbation of the backbone chain trajectory at the C-terminus (Figure 2.7) or

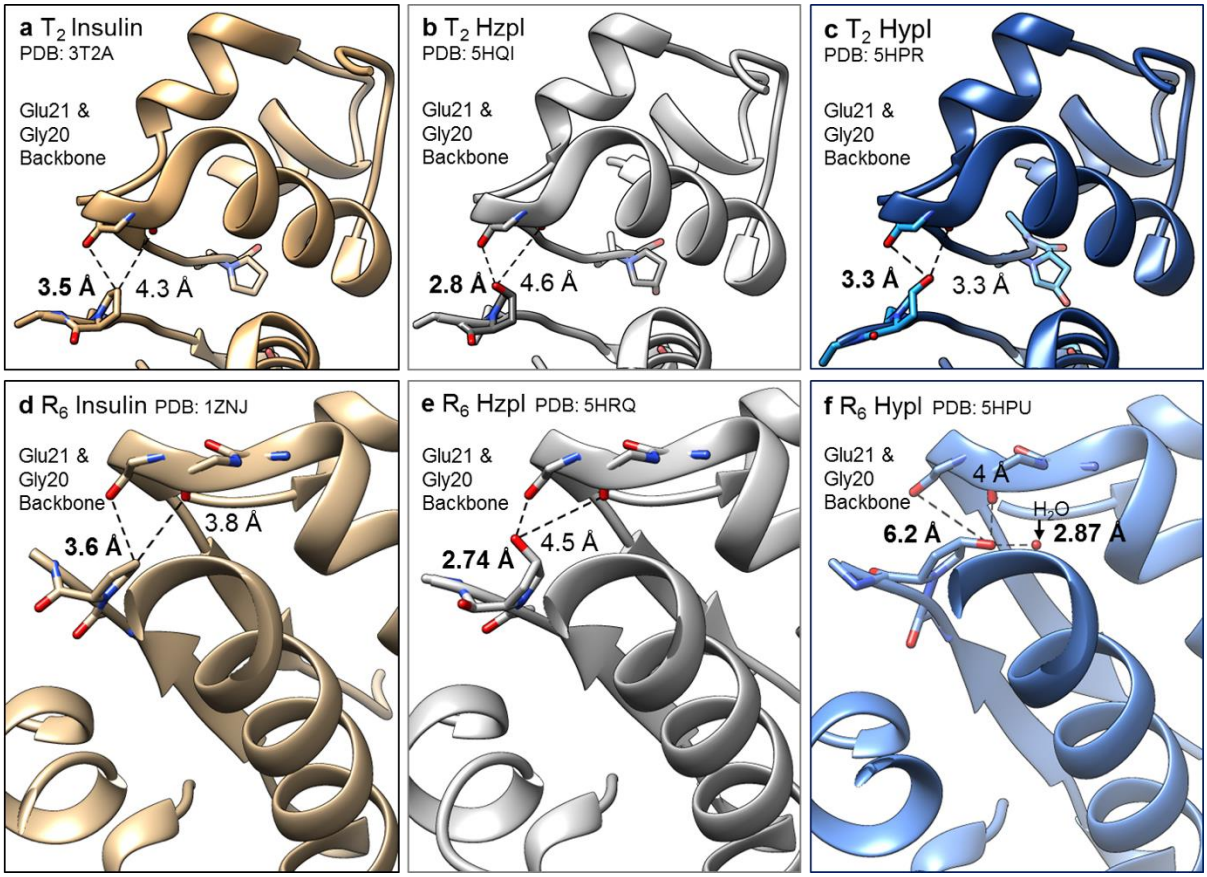
the overall insulin structure. In comparison with ProI, the backbone root-mean-square-deviation (RMSD) values of HypI and HzpI are 0.31 Å (T_2- HypI), 0.44 Å (T_2- HypI), 0.38 Å (R_6-HypI), and 0.69 Å (R_6- HypI)[49]. The most notable feature of the HzpI structures is the proximity of the hydroxyl group of Hzp to the backbone carbonyl oxygen atom of GluB21', which lies across the dimer interface (denoted by prime; Figure 2.8b,e). The inter-oxygen distances (2.8 Å in the T_2 structure, 2.7 Å in R_6), are consistent with the formation of strong hydrogen bonds between the hydroxyl group of HzpB28 and the backbone carbonyl of GluB21' in both T_2-HzpI and R_6-HzpI structures. An analogous hydrogen bond has been observed in a structure (PDB ID: 1ZEH) of R_6-AspI co-crystallized with *m*-cresol[50]; here, the phenolic ligand serves as the hydrogen-bond donor. Although the significance of this hydrogen bond has not been discussed in the literature, we suggest that it may play an important role in determining the relative stabilities of the insulin species involved in dissociation and fibrillation. In contrast to the (4*S*)-hydroxyl group of Hzp, the (4*R*)-hydroxyl of HypB28 does not contact any crystallographically resolved hydrogen bond acceptor in the T_2-structure (Figure 2.8c), and appears to bond to an ordered water molecule in the R_6-hexamer (Figure 2.8f). The absence of new hydrogen-bonding interactions is consistent with the unaltered dissociation and fibrillation kinetics of HypI.

We suggest that the hydrogen bond between Hzp and GluB21' may stabilize the dimer relative to the hexamer, or perhaps reduce the energy of the transition state for the conformational change from the R-state to T-state, and thereby increase the rate of disassembly. In addition, we propose that this hydrogen bond may prevent the *C*-terminus of the B chain from fraying and initiating fibril formation. Regardless of the mechanism,

these results demonstrate the usefulness of ncAA mutagenesis, an approach that fuses the concepts of medicinal chemistry and protein design, and paves the way to further engineer insulin (discussed further in Chapters 3-4) and other therapeutic proteins.

Figure 2.8 | Crystal structures of HzpI and HypI.

*In tan (left), wt insulins from PDB (3T2A, 1ZNJ) highlighting the distance between the γ-carbon of ProB28 and its closest neighbors, backbone carbonyl oxygen atoms of GlyB20' and GluB21 in the T_2 dimer (**a**) and R_6 hexamer (**d**) forms. In grey (middle), HzpI in the T_2 dimer (**b**) and R_6 hexamer (**e**) forms. In blue (right), HypI in the T_2 dimer (**d**) and R_6 hexamer (**g**) forms.*



Materials and Methods

Materials. All canonical amino acids and (4*R*)-hydroxy-L-proline (Hyp) were purchased from Sigma. *(4S)*-hydroxy-L-proline (Hzp) was purchased from Bachem Americas. All solutions and buffers were made using double-distilled water (ddH_2O).

Strains and plasmids. The proinsulin (PI) gene with an *N*-terminal hexa-histidine tag (6xHIS), and flanked by *Eco*R1 and *Bam*H1 cut sites was ordered as a gBlock (Integrated DNA Technologies). Both the gBlock and vector pQE80L for IPTG-inducible expression were digested with *Eco*RI and *Bam*HI. Linearized vector pQE80L was dephosphorylated by alkaline phosphatase (NEB). Ligation of the digested PI gene and linearized vector yielded plasmid pQE80PI (to produce ProI). To make plasmid pQE80PI-proS (to produce HzpI and HypI): Genomic DNA was extracted from *E. coli* strain DH10β using DNeasy Blood and Tissue Kit (Qiagen). Primers (Integrated DNA Technologies) were designed to amplify the *E. coli proS* gene, encoding prolyl-tRNA synthetase, under constitutive control of its endogenous promoter, from purified genomic DNA, and to append *Nhe*I and *Nco*I sites. The digested *proS* gene was then inserted into pQE80PI between transcription termination sites by ligation at *Nhe*I and *Nco*I restriction sites. Proline-auxotrophic *E. coli* strain CAG18515 was obtained from the Coli Genetic Stock Center at Yale University. Prototrophic *E. coli* strain BL-21 was used for rich media expression of canonical insulins (ProI, AspI). Site-directed mutagenesis of pQE80PI at B28 was performed to make plasmid pQE80PI-asp, which differs from pQE80PI by three nucleotides that specify a single amino acid mutation to aspartic acid. All genes and plasmids were confirmed by DNA sequencing.

Protein expression. Plasmids pQE80PI and pQE80PI-asp were transformed into BL21 cells and grown on ampicillin-selective agar plates. A single colony was used to inoculate 5 mL of Luria-Bertani (LB) medium and grown overnight; the resulting saturated culture was used to inoculate another 1 L of LB medium. All expression experiments were conducted at 37°C, 200 RPM in shake flasks (Fernbach 2.8 L flasks, Pyrex®). Each culture was induced with 1

mM IPTG at mid-exponential phase (OD_{600} ~0.8). For incorporation of Hyp and Hzp, pQE80PI-proS was transformed into CAG18515 cells, which were grown on ampicillin-selective agar plates. To facilitate growth, a single colony was used to inoculate 25 mL of LB medium and the culture was grown overnight prior to dilution into 1 L of 1X M9, 20 aa medium (8.5 mM NaCl, 18.7 mM NH_4Cl, 22 mM KH_2PO_4, 47.8 mM Na_2HPO_4, 0.1 mM $CaCl_2$, 1 mM $MgSO_4$, 3 mg/L $FeSO_4$, 1 µg/L of trace metals (Cu^{2+}, Mn^{2+}, Zn^{2+}, MoO_4^{2-}), 35 mg/L thiamine hydrochloride, 10 mg/L biotin, 20 mM D-glucose, 200 mg/L ampicillin with 50 mg/L of L-amino acids, each). At an appropriate cell density (OD_{600} ~0.8), the culture was subjected to a medium shift; briefly, cells were centrifuged and washed with saline prior to resuspension into 0.8 L of 1.25X M9, 19 aa (1X M9, 20 aa medium without L-proline). After cells were further incubated for 30 min to deplete intracellular L-proline, 200 mL of 5X additives (1.5 M NaCl, 2.5 mM Hyp or Hzp) was added to the culture. After another 15 min of incubation at 37°C to allow amino acid uptake prior to induction, IPTG was added to a final concentration of 1 mM. At the end of 2 h, cells were harvested by centrifugation and stored at -80°C until further use.

Cell lysis and refolding from inclusion bodies. Cells were thawed on the benchtop for 15 min prior to resuspension in lysis buffer (B-PER®, 0.5 mg/mL lysozyme, 50 U/mL benzonase nuclease). Cells were gently agitated at RT for 1 h prior to centrifugation (10 000 g, 10 min, RT); supernatant was discarded and the pellet was washed thrice: once with wash buffer (2 M urea, 20 mM Tris, 1% Triton X-100, pH 8.0) and twice with sterile ddH_2O; centrifugation followed each wash and the supernatant was discarded. The final washed pellet containing inclusion bodies (IBs, ~50% PI) was re-suspended in Ni-NTA binding buffer (8 M urea, 300

mM NaCl, 50 mM NaH_2PO_4, pH 8.0) overnight at 4°C or at RT for 2 h, both with gentle agitation. The suspension was centrifuged to remove insoluble debris; the remaining pellet was discarded and the supernatant was mixed with pre-equilibrated Ni-NTA resin (Qiagen) at RT for 1 h in order to purify PI from the IB fraction. Unbound proteins in the IB fraction were collected in the flow-through (FT), and the resin was washed with Ni-NTA wash buffer (8 M urea, 20 mM Tris base, 5 mM imidazole, pH 8.0) and Ni-NTA rinse buffer (8 M urea, 20 mM Tris base, pH 8.0) prior to stripping PI from the resin with Ni-NTA elution buffer (8 M urea, 20 mM Tris base, pH 3.0). Fractions (IBs, FT, W, elution) were collected and run under reducing conditions on SDS-PAGE (Bis/Tris gels, Novex®); elution fractions containing PI were pooled and solution pH was adjusted to 9.6 with 6 N NaOH in preparation for oxidative sulfitolysis. Oxidative sulfitolysis was performed at RT for 4 h, with the addition of sodium sulfite and sodium tetrathionate (0.2 M Na_2SO_3, 0.02 M $Na_2S_4O_6$); the reaction was quenched by 10-fold dilution with ddH_2O. To isolate PI from the quenched solution, the pH was adjusted to between 3.5 and 4.5 by adding 6 N HCl dropwise; the solution became cloudy. The solution was centrifuged (10 000 g, 10 min, RT) and supernatant discarded. The PI pellet was then re-suspended in refolding buffer (0.3 M urea, 50 mM glycine, pH 10.6) and protein concentration was estimated by the bicinchoninic acid assay (BCA assay, Pierce®). The concentration of PI was adjusted to 0.5 mg/mL. Refolding was initiated by addition of β-mercaptoethanol to a final concentration of 0.5 mM and allowed to proceed at 12°C overnight with gentle agitation (New Brunswick® shaker, 100 RPM). Post-refolding, soluble PI was harvested by adjusting the pH of the solution to 4-5 by dropwise addition of 6 N HCl and by high speed centrifugation to remove insoluble proteins. The supernatant was

adjusted to pH 8-8.5 by dropwise addition of 6 N NaOH and dialyzed against fresh PI dialysis buffer (7.5 mM sodium phosphate buffer, pH 8.0) at 4°C with five buffer changes to remove urea. The retentate (PI in dialysis buffer) was then lyophilized and subsequently stored at -80°C until further processing. Typical yields were 25-50 mg PI per L of culture (25-30 mg/L for non-canonical PI, 40-50 mg/L for canonical PI expression in rich media)

Proteolysis and chromatographic (HPLC) purification. The dry PI powder was re-dissolved in water to a final concentration of 5 mg/mL PI (final concentration of sodium phosphate buffer is 100 mM, pH 8.0). Trypsin (Sigma-Aldrich) and carboxypeptidase-B (Worthington Biochemical) were added to final concentrations of 20 U/mL and 10 U/mL, respectively to initiate proteolytic cleavage. The PI/protease solution was incubated at 37°C for 2.5 h; proteolysis was quenched by addition of 0.1% trifluoroacetic acid (TFA) and dilute HCl to adjust the pH to 4. Matured insulin was purified by reversed phase high-performance liquid chromatography (HPLC) on a C_{18} column using a gradient mobile phase of 0.1% TFA in water (solvent, A) and 0.1% TFA in acetonitrile (ACN; solvent, B). Elution was carried from 0% B to 39% B with a gradient of 0.25% B per minute during peak elution. Fractions were collected and lyophilized, and the dry powder was re-suspended into 10 mM sodium phosphate, pH 8.0. Insulin-containing fractions were verified by matrix-assisted laser desorption/ionization-mass spectrometry (MALDI-MS; Voyager MALDI-TOF, Applied Biosystems) and SDS-PAGE to ensure identify and purity. Typical yields were 5-10 mg insulin per 100 mg PI. Fractions were stored at -80°C in 10 mM phosphate buffer, pH 8.0 until further use.

Verification of Hyp and Hzp incorporation levels and maturation. A 30 µL aliquot of PI solution (8 M urea, 20 mM Tris, pH 8) was subjected to cysteine reduction and alkylation (5 mM DTT, 55°C, 20 min; 15 mM iodoacetamide, RT, 15 min, dark) prior to 10-fold dilution into 100 mM NH_4HCO_3, pH 8.0 (100 µL final volume). Peptide digestion was initiated with 0.6 µL of gluC stock solution (reconstituted at 0.5 µg/µL with ddH_2O, Promega) at 37°C for 2.5 h. The reaction was quenched by adding 10 µL of 5% TFA and immediately subjected to C_{18} ZipTip (Millipore) peptide purification and desalting according to the manufacturer's protocol. Peptides were eluted in 50% ACN, 0.1% TFA; the eluent was then diluted three-fold into matrix solution (saturated α-cyanohydroxycinnamic acid in 50% ACN, 0.1% TFA) and analyzed by mass spectrometry (Voyager MALDI-TOF, Applied Biosystems). Hyp and Hzp incorporation levels were analyzed prior to and after refolding; incorporation percentage was calculated by comparing total AUC (area under the curve, arbitrary units) of the non-canonical peak (1573 Da for the proinsulin peptide containing B28Hzp or B28Hyp, approximately 5824 Da for intact HzpI and HypI) with total AUC of its wild-type counterpart (1557 Da and 5808 Da, respectively). Maturation of HypI and HzpI was analyzed after HPLC purification. TFA (1.6 µL, 5%) was added to 15 µL mature insulin solution (10 mM phosphate buffer pH 8.0) and subjected to C_{18} ZipTip (Millipore) peptide purification and desalting per the manufacturer's protocol. MALDI-MS conditions described above were used to confirm insulin maturation.

Insulin receptor (IR) phosphorylation immunoblot. *In vitro* analysis of insulin receptor (IR) phosphorylation was performed using HEK293 cells according to a previous report[25]. Briefly, HEK293 cells were maintained in a 37°C, 5% CO_2 humidified incubator chamber

using Dulbecco's modified Eagle's medium with 4.5 g/L glucose, 2 mM L-glutamine and phenol red (DMEM, Life Technologies) supplemented with 10% fetal bovine serum (FBS, Life Technologies), 5% penicillin/streptomycin (P/S, Life Technologies). Every 3 days, at approximately 80% confluency, cells were subcultured and seeded in a 6-well plate at a cell density of $8x10^3$ cells / cm^2 (or $8x10^4$ cells per well) for 24 h prior to insulin addition. Insulins or vehicle were added directly to the medium at 200 nM (10 µL of a 50 µM solution in vehicle PBS) and incubated for 10 min prior to PBS washes to remove excess medium. HEK293 cells were lysed on-plate using IP Lysis Buffer (ThermoFisher, Pierce) with 50 U/mL benzonase nuclease (Sigma-Aldrich) for 20 min at 4°C; lysates were precipitated using ice cold acetone and re-suspended in 8 M urea, 20 mM Tris, pH 10.0. The protein concentration in the lysate was quantified by the BCA assay (ThermoFisher, Pierce) according to the manufacturer's protocol and normalized for even protein loading across lanes. Lysates were separated by SDS-PAGE (4-12% Novex Bis/Tris SDS-PAGE gels, Life Technologies) in duplicate and transferred to a nitrocellulose membrane (Hybond ECL, GE Healthcare) using a wet transfer system. The membrane was blocked at RT in 5% nonfat milk in Tris-buffer saline with 0.1% Tween 20 (TBS/Tween) and washed with TBS/Tween prior to blotting with antibodies. Primary antibodies for insulin receptor, phosphorylated insulin receptor (from Cell Signaling Technologies) and β-actin (as loading control, from Invitrogen) were added at 1:1000 dilution in TBS/Tween with gentle agitation either at RT for 4 h or overnight at 4°C. Blots were washed and secondary antibodies (Invitrogen) were added at 1:2000 dilution in TBS/Tween. Blots were washed again prior to fluorescence imaging on a Typhoon Trio (GE Healthcare).

Reduction of blood glucose in diabetic animals. NODscid (NOD.CB17-*Prkdc*^{scid}/J) mice were obtained from Jax Mice (Bar Harbor, Maine). Mice were maintained under specific pathogen-free conditions, and experiments were conducted according to procedures approved by the Institutional Animal Care and Use Committee at the City of Hope. Adult (8- to 12-week-old) male NODscid mice were injected intraperitoneally (50 mg/kg/day for 3 consecutive days) with freshly prepared streptozotocin (STZ) in 0.05 M 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 (defined as >200 mg/dL), measured by using a glucomonitor (FreeStyle; Abbott Diabetes Care, Alameda, CA) in blood (10 µL) sampled from the lateral tail vein. Insulin analogs concentrations were determined from A_{280} measurements using a molar extinction coefficient of 6080 M^{-1} cm^{-1} and diluted to 100 µg/mL into a formulation buffer according to a previous report[47]. Insulin analogs in solution were injected subcutaneously at the scruff and blood glucose was measured at the indicated time points.

Circular Dichroism. Spectra were collected in a 1 cm quartz cuvette at an insulin concentration of 60 µM in 50 mM sodium phosphate buffer pH 8.0. Data were collected from 185 nm to 250 nm, with step size of 0.25 nm and averaging time of 1 s on a Model 410 Aviv Circular Dichroism Spectrophotometer; spectra were averaged over 3 repeat scans. A reference buffer spectrum was subtracted from the sample spectra for conversion to mean residue ellipticity.

Analytical Ultracentrifugation. Sedimentation velocity (SV) and sedimentation equilibrium (SE) experiments were carried out on an XL-1 AUC (BeckmanCoulter). SV experiments were conducted with insulin samples dialyzed against 50 mM Tris, 0.1 mM EDTA, pH 8.0, which

also served as the reference buffer. Two sector cells with sapphire windows were filled with sample and reference buffer. These cells were centrifuged at 50,000 RPM with absorbance data collected at 280 nm, or for concentrations above 1 mg/mL, 281 nm or 287 nm. SV data were analyzed in SEDFIT with the c(s) algorithm for a continuous distribution[51]. Buffer density and viscosity were calculated from SEDNTERP; the partial specific insulin volume used was 0.735[52]. SE experiments were conducted with insulin samples dialyzed against 50 mM Tris, 0.1 mM EDTA, pH 8.0, which also served as the reference buffer. Two sector cells with sapphire windows were filled with sample and reference buffer and centrifuged at 15,000, 24,000, 36,000, and 50,000 RPM with absorbance data collected at 280 nm. Equilibrium was ascertained by analysis in SEDFIT and non-equilibrated scan speeds were excluded from data analysis. SE and SV data from multiple concentrations were fitted to a monomer-dimer-hexamer reversible self-association model in SEDPHAT with best model chosen by inspection of residuals as well as critical χ value deviation[53]. Radial dependent baselines were computationally determined using TI noise. Figures were generated using GUSSI[54].

Hexamer Dissociation Assay. Insulins were quantified by both UV absorbance (NanoDrop Lite, ThermoFisher) and BCA assay, and normalized to 125 µM insulin prior to dialysis against 50 mM Tris/perchlorate, 25 µM zinc sulfate, pH 8.0 overnight at 4°C using a D-tube dialyzer (Millipore Corp.) with MWCO of 3.5 kDa. Aliquots of dialyzed insulin solution were mixed with phenol to yield samples of the following composition: 100 µM insulin, 20 µM zinc sulfate, 100 mM phenol. Dissociation was initiated by addition of terpyridine (Sigma-Aldrich) to a final concentration of 0.3 mM from a 0.75 mM stock solution. A Varioskan

multimode plate reader (Thermo Scientific) was used to monitor absorbance at 334 nm. Kinetic runs were done at least in triplicate, and the data were fit to a mono-exponential function using Origin software. Post assay insulin samples were pooled and sample quality was determined by SDS-PAGE.

Fibrillation Assay. Insulin samples (60 µM in 10 mM phosphate, pH 8.0) were centrifuged at 22 000 g for 1 h immediately after addition of thioflavin T (ThT) (EMD Millipore) to a final concentration of 1 µM. Samples were continuously shaken at 960 RPM on a Varioskan multimode plate reader at 37°C, and fluorescence readings were recorded every 15 min for 48 h (excitation 444 nm, emission 485 nm). Assays were run in quadruplicate, in volumes of 200 µL in sealed (Perkin-Elmer), black, clear-bottom 96 well plates (Grenier BioOne).

Transmission electron microscopy. Insulin samples (60 µM in 10 mM phosphate, pH 8.0) were continuously agitated in microfuge tubes at 42°C, 960 RPM for 48 h in a ThermoMixer to obtain fibrils. Samples were stained onto 200 mesh copper grids (formar/carbon coated, plasma cleaned) with 1% uranyl acetate. Imaging was done by Alasdair McDowell at the Beckman Institute's Center for Transmission Electron Microscopy on a Tecnai T12 LaB6 120 eV Transmission Electron Microscope.

Crystallographic Studies. Insulin crystals were obtained from sitting drop trays set using a Mosquito robot (TTP Labtech). Drops were set by mixing 0.4 µL insulin solution with 0.4 µL well solution. Well solution conditions were as follows: 462.5 mM sodium citrate, 100 mM HEPES, pH 8.25 for 5HQI; 300 mM Tris, 0.5 mM zinc acetate, 8.5% acetone, 0.5 M sodium citrate pH 8.0 for 5HPR; 300 mM Tris, 17 mM zinc acetate, 1% phenol, 7.5% acetone, 2.675 M sodium citrate pH 8.0 for 5HRQ; 300 mM Tris, 17 mM zinc acetate, 1% phenol, 7.5%

acetone, 1.95 M sodium citrate pH 8.0 for 5HPU. Cells were cryoprotected in a mother liquor containing 30% glycerol prior to looping and flash freezing in liquid nitrogen. Data were collected at SSRL beamline BL12-2 using a DECTRIS PILATUS 6M pixel detector. Initial indexing and scaling was performed with XDS; for some structures, data were re-scaled in alternative space groups using Aimless[55]. Initial phases were generated by molecular replacement in PHASER with 3T2A (5HQI and 5HPR) or 1EV3 (5HRQ and 5HPU)[56]. Structure refinement was carried out in Coot and Refmac5[57,58]. Data were deposited in the PDB with the following codes: 5HQI (T_2-HzpI), 5HPR (T_2-HypI), 5HRQ (R_6-HzpI), 5HPU (R_6-HypI). All distances and contacts were computed using crystallography software (UCSF Chimera).

References

1. Brange, J. et al. Monomeric insulins obtained by protein engineering and their medical implications. *Nature* **333**(6174): 679-682 (1988).
2. Ciszak, E. et al. Role of C-terminal B-chain residues in insulin assembly: the structure of hexameric LysB28ProB29-human insulin. *Structure* **3**(6): 615-22 (1995).
3. Owens, D.R. & Vora, J. Insulin aspart. *Expert Opin Drug Metab Toxicol* **2**(5): 793-804 (2006).
4. Dodson, G. & Steiner, D. The role of assembly in insulin's biosynthesis. *Curr Opin Struct Biol* **8**(2): 189-94 (1998).
5. Rutter, Guy A., Pullen, Timothy J., Hodson, David J. & Martinez-Sanchez, A. Pancreatic β-cell identity, glucose sensing and the control of insulin secretion. *Biochemical Journal* **466**(2): 203 (2015).
6. Pu, Y., Lee, S., Samuels, D.C., Watson, L.T. & Cao, Y. The effect of unhealthy β-cells on insulin secretion in pancreatic islets. *BMC Medical Genomics* **6**(Suppl 3): S6-S6 (2013).
7. Rorsman, P. & Braun, M. Regulation of insulin secretion in human pancreatic islets. *Annu Rev. Physiol* **75**: 155-79 (2013).
8. Baker, E.N. et al. The Structure of 2Zn Pig Insulin Crystals at 1.5 A Resolution. *Philos Trans R Soc Lond B Biol Sci.* **319**(1195): 369 (1988).
9. Mark, A.E., Nichol, L.W. & Jeffrey, P.D. The self-association of zinc-free bovine insulin. *Biophys. Chem.* **27**(2): 103-117 (1987).
10. Hansen, J.F. The self-association of zinc-free human insulin and insulin analogue B13-glutamine. *Biophys. Chem.* **39**(1): 107-110 (1991).

11. Brange, J., Andersen, L., Laursen, E.D., Meyn, G. & Rasmussen, E. Toward understanding insulin fibrillation. *J Pharm Sci* **86**(5): 517-525 (1997).
12. Zoete, V., Meuwly, M. & Karplus, M. A Comparison of the Dynamic Behavior of Monomeric and Dimeric Insulin Shows Structural Rearrangements in the Active Monomer. *J Mol Biol* **342**(3): 913-929 (2004).
13. Brems, D.N. et al. Altering the association properties of insulin by amino acid replacement. *Protein Eng.* **5**(6): 527-533 (1992).
14. Cowell, S.M., Lee, Y.S., Cain, J.P. & Hruby, V.J. Exploring Ramachandran and Chi Space: Conformationally Constrained Amino Acids and Peptides in the Design of Bioactive Polypeptide Ligands. *Current Medicinal Chemistry* **11**(21): 2785-98 (2004).
15. Nishi, Y. et al. Different effects of 4-hydroxyproline and 4-fluoroproline on the stability of collagen triple helix. *Biochemistry* **44**(16): 6034-6042 (2005).
16. Shoulders, M.D., Kotch, F.W., Choudhary, A., Guzei, I.A. & Raines, R.T. The aberrance of the 4*S* diastereomer of 4-hydroxyproline. *JACS* **132**(31): 10857-10865 (2010).
17. Kuemin, M. et al. Tuning the cis/trans Conformer Ratio of Xaa–Pro Amide Bonds by Intramolecular Hydrogen Bonds: The Effect on PPII Helix Stability. *Angew Chem Int Ed Engl* **49**(36): 6324-6327 (2010).
18. Erdmann, R.S. & Wennemers, H. Importance of Ring Puckering versus Interstrand Hydrogen Bonds for the Conformational Stability of Collagen. *Angew Chem Int Ed Engl* **50**(30): 6835-6838 (2011).
19. Bretscher, L.E., Jenkins, C.L., Taylor, K.M., DeRider, M.L. & Raines, R.T. Conformational stability of collagen relies on a stereoelectronic effect. *JACS* **123**(4): 777-778 (2001).
20. Kemmler, W., Peterson, J.D. & Steiner, D.F. Studies on the conversion of proinsulin to insulin. *J. Biol. Chem.* **246**(22): 6786-6791 (1971).
21. 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.* **151**(4): 350-356 (2011).
22. 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* **5**(7): 928-936 (2004).
23. 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* (0021-9193 (Print))(1986).
24. Pandyarajan, V. et al. Biophysical optimization of a therapeutic protein by non-standard mutagenesis: studies of an iodo-insulin derivative. *J. Biol. Chem.* **289**: 23367-23381 (2014).
25. Wharton, J., Meshulamy, T., Vallega, G. & Pilch, P. Dissociation of insulin receptor expression and signaling from Caveolin-1 expression. *J. Biol. Chem.* **280**(14): 13483-13486 (2005).
26. Menting, J.G. et al. Protective hinge in insulin opens to enable its receptor engagement. *Proc. Natl. Acad. Sci. U. S. A.* **111**(33): E3395-404 (2014).
27. Antolikova, E. et al. Non-equivalent role of inter- and intramolecular hydrogen bonds in the insulin dimer interface. *J. Biol. Chem.* **286**(42): 36968-77 (2011).

28. Attri, A.K., Fernández, C. & Minton, A.P. pH-dependent self-association of zinc-free Insulin characterized by concentration-gradient static light scattering. *Biophys. Chem.* **148**(1-3): 28-33 (2010).
29. Zhao, H. & Schuck, P. Combining biophysical methods for the analysis of protein complex stoichiometry and affinity in SEDPHAT. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **71**(Pt 1): 3-14 (2015).
30. Brown, P.H. & Schuck, P. Macromolecular size-and-shape distributions by sedimentation velocity analytical ultracentrifugation. *Biophys. J.* **90**(12): 4651-61 (2006).
31. Pandyarajan, V. & Weiss, M.A. Design of non-standard insulin analogs for the treatment of diabetes mellitus. *Curr. Diab. Rep.* **12**(6): 697-704 (2012).
32. Birnbaum, D.T., Kilcomons, M.A., DeFelippis, M.R. & Beals, J.M. Assembly and dissociation of human insulin and LysB28ProB29-insulin hexamers: a comparison study. *Pharm. Res.* **14**(1): 25-36 (1997).
33. Rahuel-Clermont, S., French, C.A., Kaarsholm, N.C. & Dunn, M.F. Mechanisms of stabilization of the insulin hexamer through allosteric ligand interactions. *Biochemistry* **36**(19): 5837-5845 (1997).
34. Dische, F. et al. Insulin as an amyloid-fibril protein at sites of repeated insulin injections in a diabetic patient. *Diabetologia* **31**(3): 158-161 (1988).
35. Brange, J., Dodson, G.G., Edwards, D.J., Holden, P.H. & Whittingham, J.L. A model of insulin fibrils derived from the x-ray crystal structure of a monomeric insulin (despentapeptide insulin). *Proteins: Structure, Function, and Genetics* **27**(4): 507-516 (1997).
36. Nielsen, L., Frokjaer, S., Brange, J., Uversky, V.N. & Fink, A.L. Probing the mechanism of insulin fibril formation with insulin mutants. *Biochemistry* **40**(28): 8397-8409 (2001).
37. Whittingham, J.L. et al. Insulin at pH 2: Structural analysis of the conditions promoting insulin fibre formation. *J Mol Biol* **318**(2): 479-490 (2002).
38. Ahmad, A., Millett, I.S., Doniach, S., Uversky, V.N. & Fink, A.L. Partially folded intermediates in insulin fibrillation. *Biochemistry* **42**(39): 11404-11416 (2003).
39. Hua, Q.-X. & Weiss, M.A. Mechanism of insulin fibrillation: The structure of insulin under amyloidogenic conditions resembles a protein-folding intermediate. *J. Biol. Chem.* **279**(20): 21449-21460 (2004).
40. Huang, K., Maiti, N.C., Phillips, N.B., Carey, P.R. & Weiss, M.A. Structure-specific effects of protein topology on cross-β assembly: studies of insulin fibrillation. *Biochemistry* **45**(34): 10278-10293 (2006).
41. Hong, D.-P., Ahmad, A. & Fink, A.L. Fibrillation of human insulin A and B chains. *Biochemistry* **45**(30): 9342-9353 (2006).
42. Ivanova, M.I., Sievers, S.A., Sawaya, M.R., Wall, J.S. & Eisenberg, D. Molecular basis for insulin fibril assembly. *Proc. Natl. Acad. Sci. U. S. A.* **106**(45): 18990-18995 (2009).
43. Nayak, A., Sorci, M., Krueger, S. & Belfort, G. A universal pathway for amyloid nucleus and precursor formation for insulin. *Proteins: Structure, Function, and Bioinformatics* **74**(3): 556-565 (2009).

44. Babenko, V. & Dzwolak, W. Amino acid sequence determinants in self-assembly of insulin chiral amyloid superstructures: Role of C-terminus of B-chain in association of fibrils. *FEBS Letters* **587**(6): 625-630 (2013).
45. Noormägi, A., Valmsen, K., Tõugu, V. & Palumaa, P. Insulin fibrillization at acidic and physiological pH values is controlled by different molecular mechanisms. *The Protein Journal*: 1-6 (2015).
46. Chatani, E. et al. Early aggregation preceding the nucleation of insulin amyloid fibrils as monitored by small angle X-ray scattering. *Scientific Reports* **5**: 15485 (2015).
47. Pandyarajan, V. et al. Aromatic anchor at an invariant hormone-receptor interface: function of insulin residue B24 with application to protein design. *J. Biol. Chem.* **289**(50): 34709-27 (2014).
48. Vinther, T.N. et al. Novel covalently linked insulin dimer engineered to investigate the function of insulin dimerization. *PLoS ONE* **7**(2): e30882 (2012).
49. Marshall, H., Venkat, M., Seng, N.S., Cahn, J. & Juers, D.H. The use of trimethylamine N-oxide as a primary precipitating agent and related methylamine osmolytes as cryoprotective agents for macromolecular crystallography. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **68**(Pt 1): 69-81 (2012).
50. Smith, G.D., Ciszak, E., Magrum, L.A., Pangborn, W.A. & Blessing, R.H. R6 hexameric insulin complexed with *m*-cresol or resorcinol. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **56**(12): 1541-1548 (2000).
51. Schuck, P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys. J.* **78**(3): 1606-1619 (2000).
52. Laue, T.M., Shah, B.D., Ridgeway, T.M. & Pelletier, S.L. *Analytical ultracentrifugation in biochemistry and polymer science*, (Royal Society of Chemistry, Cambridge [England], 1992).
53. Vistica, J. et al. Sedimentation equilibrium analysis of protein interactions with global implicit mass conservation constraints and systematic noise decomposition. *Anal. Biochem* **326**(2): 234-256 (2004).
54. Brautigam, C.A. Chapter five - calculations and publication-quality illustrations for analytical ultracentrifugation data. in *Methods Enzymol.*, Vol. Volume 562 (ed. James, L.C.) 109-133 (Academic Press, 2015).
55. Winn, M.D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **67**(Pt 4): 235-242 (2011).
56. McCoy, A.J. et al. Phaser crystallographic software. *J Appl Crystallogr.* **40**(Pt 4): 658-674 (2007).
57. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **66**(Pt 4): 486-501 (2010).
58. Murshudov, G.N., Vagin, A.A. & Dodson, E.J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **53**(3): 240-255 (1997).

Acknowledgements

We thank J. T. Kaiser and P. Nikolovski of the Molecular Observatory at Caltech, and S. Russi and the scientific staff of Beamline 12-2 at the Stanford Synchrotron Radiation Laboratory for assistance. We thank S. Virgil of the Chemical Catalysis Center, M. Shahgholi of the Mass Spectrometry Facility, and A. McDowall at the Center for Transmission Electron Microscopy at Caltech for their assistance. We thank W. Glenn, A. Madhavi, and T. Hoeg-Jensen for discussions, and J. Cahn for crystallography assistance. We thank T. Ku, J. Lebon, and J. Rawson at the City of Hope for performing insulin activity assays *in vivo*.

This work was done in collaboration with Seth Lieblich. S.L. performed experiments for insulin maturation and HPLC purification, sample preparation for obtaining circular dichroism spectra and *in vivo* mouse assays, analytical ultracentrifugation of insulin to obtain K_D values, and solving crystal structures of insulin.

Portions of this chapter were adapted from a manuscript (Hydroxylation of insulin at position B28 accelerates hexamer dissociation and delays fibrillation; Seth A. Lieblich, Katharine Y. Fang, Jackson K. B. Cahn, Jeffrey Rawson, Jeanne LeBon, H. Teresa Ku, & David A. Tirrell) submitted for publication.

Chapter 3 – Understanding the effects of fluorination of insulin at position B28

Abstract

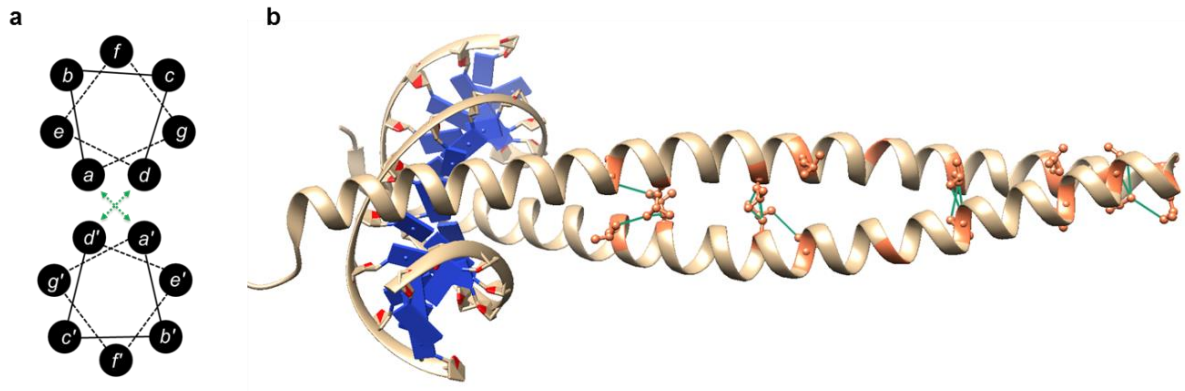
Mutating the proline residue at position 28 (ProB28) of the insulin B-chain disrupts the inter-subunit interface in the hexamer, enhances the rate of disassembly, and has served as the basis for producing rapid-acting insulins (RAIs); however, previous mutations here have not been shown to delay the onset of fibrillation[1,2]. The fluoroprolines, (*2S,4S*)-4-fluoro-L-proline (Fzp), (*2S,4R*)-4-fluoro-L-proline (Fyp), and (4,4)-difluoro-L-proline (dFp), when incorporated at proline positions in several proteins, have resulted in stabilizing effects[3-7]. Here we show that replacement of ProB28 by fluoroprolines yields forms of insulin with unique properties. Fzp-insulin (FzpI) has a similar hexamer dissociation rate and a marginally delayed lag time to fibril formation compared to the wild-type protein (ProI), while Fyp-insulin (FypI) dissociates like known RAIs, but forms fibrils more readily than ProI. dFpI-insulin (dFpI) has intermediate dissociation rate (between ProI and RAIs) but is, surprisingly, also the most stable fluoroinsulin. Our results demonstrate that the biophysical behavior of insulin is sensitive to fluorination at ProB28, and crystal structures suggest that CH/π interactions between ProB28 and TyrB26 may play a larger role in insulin biophysics than what has been previously reported in the literature.

Introduction

Fluorine and Fluorinated Proteins. Fluorine has a long and storied use in medicinal chemistry[8] because of its small size (minimal structural perturbations) and hydrophobicity. Fluorine is highly electronegative, rendering the carbon-fluorine (C-F) bond extremely strong and polarized[9], which can influence lipophilicity (to cross cell membranes), tune pKa, and solubility[10]. In addition, the fluorination of small molecules has generally led to increased potency (e.g., Fludrocortisone is more potent than the parent compound Hydrocortisone)[8]. Fluorine does not appear[11] in natural proteins, although there are known organofluoro molecules that have been found in a handful of species[12]. Protein engineers have taken advantage of fluorine's expanded hydrophobic surface area and electronegative effects[8,9,13-15] through replacement of interacting residues in leucine zipper (Figure 3.1) and collagen helices with fluorinated non-canonical amino acids (ncAAs).

Figure 3.1 | Leucine zipper used in fluorination studies.

***a**, Consensus motif for coiled coils in helical wheel representation. **b,** Parallel coiled-coil structure of leucine zipper GCN4 (PDB: 1GD2) with leucines colored (coral) and side chain represented by ball-and-stick. Van der Waals (vdW) contacts between leucines at the core of the helices are depicted with the dark green lines.*



Fluorination of leucines. Interactions between leucine residues drive the self-association[16] of coiled coils. Leucines at the *d* positions (Figure 3.1a) are critical to dimerization[17] because they form hydrophobic van der Waals (vdW) contacts with one another (Figure 3.1b). Several studies[11,18] on leucine zippers have detailed the stabilizing effects of fluorinated surfaces. It has been theorized that fluorination largely retains the canonical hydrophobic character, while allowing the slightly larger and more hydrophobic fluorinated groups to pack densely, which effectively stabilizes the helical structure of coiled coils[19].

Fluoroproline. Several studies have investigated the effects of fluoroprolines in proteins[5,6,20]. The natural helical structure of collagen relies on both hydrogen bonding interactions and inductive effects from the hydroxyl group of (*2S,4R*)-4-hydroxy-L-proline (Hyp)[21]. Substitution of Hyp with (*2S,4R*)-4-fluoro-L-proline (Fyp) in collagen results in a hyper-thermostable protein[3] even though Fyp is unable to form hydrogen bonding interactions; however, the fluorine of Fyp is more electronegative than the oxygen from the hydroxyl group of Hyp. Therefore, from these experimental results, it was concluded that collagen stability is more reliant on inductive effects than hydrogen bonding interactions[22,23]. However, the information obtained from a simple structure, such as collagen, is not readily applicable to globular proteins, which typically contain higher order structures with complex interactions between residues.

Folding studies on fluorinated globular proteins. Cis-trans isomerization of the peptide bond between any residue X and Pro (X-Pro) is known to govern rates of protein folding[24]. This rate-determining folding step has been experimentally determined for bovine pancreatic ribonuclease A[25], ubiquitin[5] and *E. coli* thioredoxin[6,26]; all three of these proteins require at

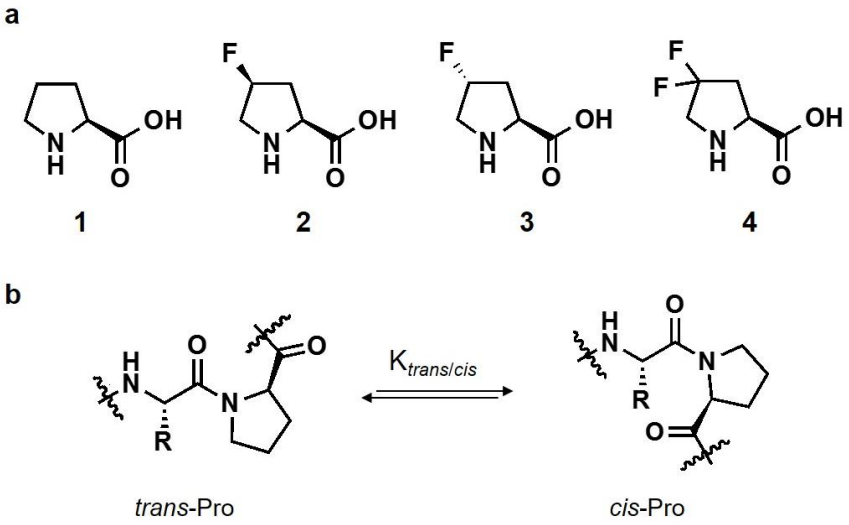
least one *cis* X-Pro to proceed to its correctly folded form. Fluoroprolines (Figure 3.2) have different *cis-trans* equilibrium constants; in model peptide studies, (*2S,4S*)-4-fluoro-L-proline (Fzp) prefers the *cis* conformation approximately 3 times more than Pro[13]. For example, replacement of the *cis*-Pro in thioredoxin by Fzp leads to an 8-fold increase in the protein folding rate[26].

Stability studies on fluorinated β2-*microglobulin.* The *cis-trans* isomerization of the proline peptide bond (Figure 3.2b) in proteins is dynamic and has consequences *in vivo*. Studies on a neurotransmitter channel (5-hydroxytrptamine type 3 receptor) found that its gatekeeper residue (for ion transport) was a critical proline, which when replaced with proline analogs of differing *cis-trans* isomerization preferences can be fine-tuned for electrophysiology[27]. It has also been shown in several proteins that protein aggregation rates are also directly correlated with *cis-trans* isomerism at critical proline residues[28-31]. For example, human β2-microglobulin (β2m), a component of the class I major histocompatibility complex (MHC I)[32], is stable under physiological conditions, but readily aggregates in patients with kidney disease who are undergoing long-term hemodialysis[33]. Due to its role in the immune system, β2m has been studied extensively[34-36] and NMR spectroscopy of both fibrillar and crystalline β2m states identified proline at position 32 (Pro32)[37] as the "switch" for amyloid formation. In soluble or crystalline form, β2m contains a *cis*-Pro32, while the aggregated state contains Pro32 in primarily *trans* peptide form. Fluoroprolines, with varying *cis-trans* isomerization equilibrium constants ($K_{trans/cis}$ ranges from 2.5 to 6.7, compared to $K_{trans/cis}$ of 4.7 for Pro)[12,38], were site-specifically incorporated into β2m at Pro32 through semi-synthesis. Unsurprisingly, Fzp, which of the fluoroprolines has the greatest preference for

the *cis*-amide bond ($K_{trans/cis}$ of 2.5), when incorporated into β2m showed the greatest propensity to deter aggregation[39]. Notably, the effects of *cis-trans* isomerization and stability by fluorination were not decoupled in these or follow-up studies[40].

Figure 3.2 | Proline analogs for fluorination study.

a*, Chemical structures for compound* **1***: L-proline; Compound* **2***: (2S,4S)-4-fluoro-L-proline (Fzp); Compound* **3***: (2S, 4R)-4-fluoro-L-proline (Fyp); Compound* **4***: (4,4)-difluoro-L-proline (dFp).* ***b,*** *Chemical depiction of cis-trans isomerization of the peptide amide bond containing proline.*



Halogen-containing insulins. Protein amyloidogenesis has been implicated in several human diseases[41], e.g., amyloid precursor protein (APP) in Alzheimer's Disease (AD) and insulin in insulin-derived amyloidogenesis. Recent literature[42] has presented biophysical characterization of an iodinated-insulin lispro variant in an effort to stabilize insulin and prevent fibril formation. Insulin lispro (LysI) was the first RAI to be engineered[43] and made through mutating ProB28 and LysB29 to LysB28 and ProB29, respectively; these mutations increased LysI's dissociation kinetics but did not result in any additional stability. It was speculated that iodination of tyrosine at position 26 of the B chain (TyrB26) to produce iodo-TyrB26 insulin lispro (iodo-LysI) could keep the beneficial rapid acting nature of LysI,

but also enhance its stability with the placement of a halogenated compound within the dimer interface[42]. Fibrillation and hexamer dissociation kinetic experiments found that iodo-TyrB26 was marginally able to resist fibril formation but the iodo-LysI hexamer dissociates at a rate similar to wild-type insulin (ProI), effectively losing its RAI status, and retains about 70% of its biological activity. Nonetheless, iodo-LysI is one of the first demonstrations in the literature where non-canonical amino acids (ncAAs) are used to modulate the biophysical properties of insulin.

Here, we propose to use fluoroprolines to modulate the biophysical properties of insulin. The small size of fluorine should be less of a perturbation than iodine; in addition, substitution at B28 is not expected to interfere with biological activity because of its distance from important binding residues[44]. Fluoroprolines have also been studied extensively in other contexts (e.g. fluorescent proteins[7,45], collagen[3,13,23,46,47], DNA polymerase[48]), which can be useful in elucidating the effects of fluorine on insulin. We hope the analysis of fluoroinsulins will allow us to further understand the biophysical behavior of insulin and as an addition to the insights gained from the hydroxyinsulins in Chapter 2, perhaps provide general guidelines for future insulin engineering.

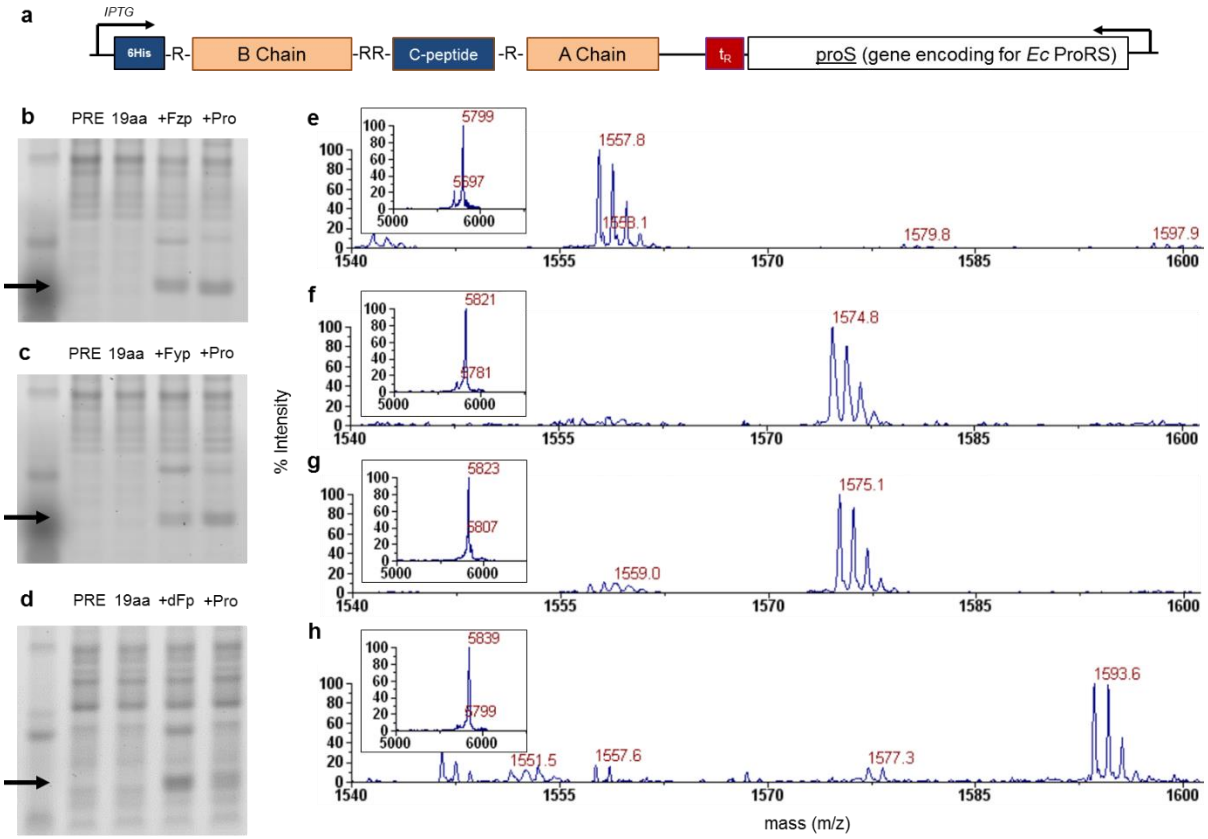
Results and Discussion

Recombinant production of biologically active FzpI, FypI and dFpI. It has been reported[49] that Fzp and Fyp can be incorporated into newly synthesized proteins in *E. coli* using the endogenous translational machinery, and dFp can be incorporated with an overexpressed prolyl-tRNA synthetase (ProRS) under osmotic stress conditions (NaCl concentration at 300 mM). Expression of proinsulin (PI) was confirmed via polyacrylamide gel electrophoresis

(Figure 3.3 b-d). Incorporation levels were assessed using matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS; Figure 3.3 e-h) and found to be approximately 90% by comparing areas of the peaks corresponding to the fluorinated or difluorinated peptides at m/z = 1575 Da or 1594 Da, respectively, and the wild-type peptide at m/z = 1558 Da.

Figure 3.3 | Insulin expression and incorporation of fluoroprolines

a, *Plasmid pQE80PI-proS containing gene construct proinsulin under inducible lac promoter and an overexpressed Ec ProRS for bacterial expression.* ***b-d,*** *SDS-PAGE of cell lysates with lanes labeled for pre-induction (PRE) and post-induction in minimal media supplemented with either nothing (19aa), Fzp (**a**), Fyp (**b**) or dFp (**c**), or Pro.* ***e-h:*** *MALDI-MS traces of isolated proinsulin peptide fragment ^{46}RGFFYT**P**KTRRE^{57} obtained by gluC digestion. Peptide fragment masses corresponds to either wild type mass 1558 Da (**e**) or shifted masses 1575 Da if Fzp (**f**) or Fyp (**g**), and 1594 Da if dFp (**h**) is incorporated. Inset is whole protein MALDI-MS (instrument error of ~10 Da, 1-2%) with observed masses: 5799 Da (**e**, ProI), 5822 Da (**f-g**, FzpI and FypI), and 5839 (**h**, dFpI). All MALDI-MS spectra contain ion counts >10^3*



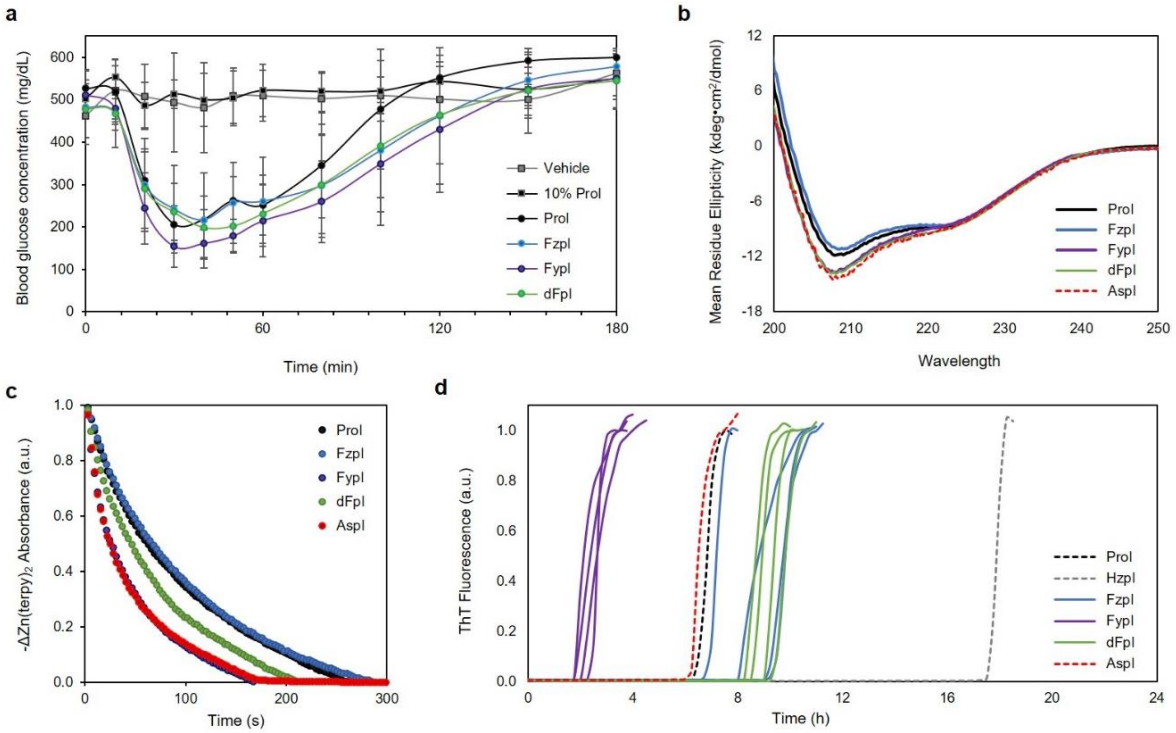
Refolding, and insulin maturation and biological activity. Inclusion bodies containing PI were extracted and resolubilized in denaturing buffer prior to purification with Ni-NTA affinity chromatography. Elution fractions containing PI were combined, pH-adjusted, and subjected to oxidative sulfitolysis, and refolded. The refolded protein was isolated by shifting to the solution to acidic conditions and then dialyzed, lyophilized and re-suspended in water for proteolytic cleavage (with trypsin and carboxypeptidase B) and liquid chromatography (HPLC) purification. Post-HPLC fractions containing mature insulin were lyophilized, re-suspended in phosphate buffer prior to verification by whole protein MALDI-MS (Figure 3.3e-h, insets) and polyacrylamide gel electrophoresis. PI and insulin yields were similar to the hydroxyinsulins described in Chapter 2. Insulin fractions were stored at -80°C and thawed on ice prior to use.

We expected our recombinant insulins to be biologically active because the modified residue B28 or nearby residues (B27, B29-B30) are not involved in binding to the insulin receptor (IR)[50]. Recombinant insulins formulated under pharmaceutical conditions were injected into streptozocin (STZ)-induced diabetic mice and able to lower blood glucose levels within 30 minutes (Figure 3.4a).

Incorporation of fluoro-prolines alters insulin dimerization. At 60 µM, in absence of Zn^{2+} and phenol, both FypI and dFp appears to be monomeric (with CD spectra nearly identical to that of AspI; Figure 3.4b) while the spectrum of FzpI suggests a dimeric insulin (with CD spectra nearly identical to that of ProI; Figure 3.4b).

Figure 3.4 | Fluorination at ProB28 does not affect biological activity but alters insulin dimerization, dissociation kinetics, and stability

a, *Reduction of blood glucose following subcutaneous injection of 35 µg/kg insulins into streptozocin-induced diabetic mice. Glucose levels were measured post-injection via tail vein sampling. ProI, 10% ProI, FzpI, FypI, dFpI or vector were formulated as described[42]. Error bars denote one standard deviation (n ≥ 3).* ***b,*** *Far UV CD spectra collected on 60 µM insulins in 10 mM phosphate buffer, pH 8.0 at 25°C.* ***c,*** *Insulin hexamer dissociation following sequestration of Zn^{2+} by terpyridine. Zn^{2+}-(terpy) signal was monitored at 334 nm and fitted to a mono-exponential decay. FzpI, FypI and dFpI contain 10% ProI. The curves for the pairs: FypI and AspI, FzpI and ProI, are indistinguishable in this plot.* ***d,*** *Representative fibrillation curves for 60 µM insulins (37°C, 960 RPM; n=4). Insulin fibrils were detected by the rise in Thioflavin T (ThT) fluorescence that occurs upon binding to fibrillar aggregates. Median curves are shown for ProI, AspI and HzpI.*

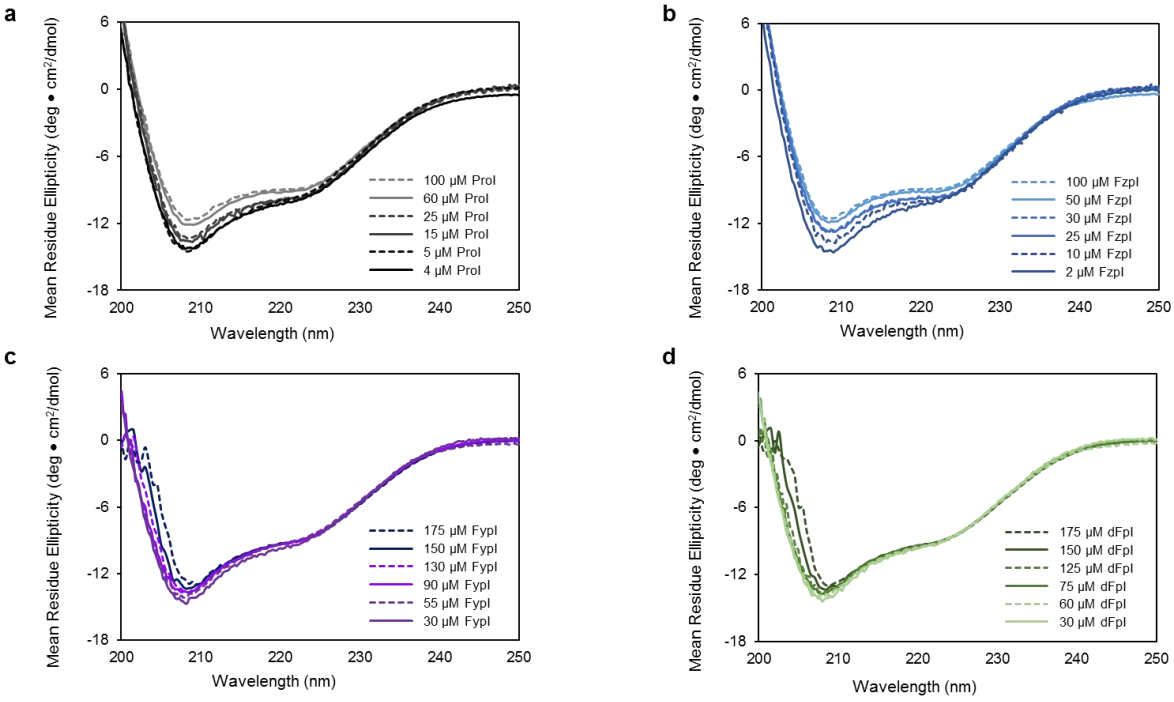


To obtain estimates for the insulin dimer's dissociation constant (K_D)[51], we obtained CD spectra at varying concentrations (Figure 3.5). FzpI has a K_D of approximately 2 µM, which is on the same order of magnitude as ProI, indicating that the *(4S)*-fluoro group of FzpB28 does not substantially affect dimerization. Both FypI and dFpI have a K_D of approximately 150 µM, indicating a perturbed dimer interface and are more monomeric than FzpI and ProI. The stereochemical trend for dimerization correlate for the mono-

fluoroinsulins and hydroxyinsulins: *(4S)*-substituents have little effect on the dimerization constant (K_D ~10 µM), while the *(4R)*-substituents results in a significantly perturbed dimer interface (K_D >100 µM).

Figure 3.5 | Far UV CD spectra of fluoroinsulins at varying concentrations.

*ProI (**a**), FzpI (**b**), FypI (**c**), and dFpI (**d**) in 10 mM phosphate buffer, pH 8.0 at 25°C.*



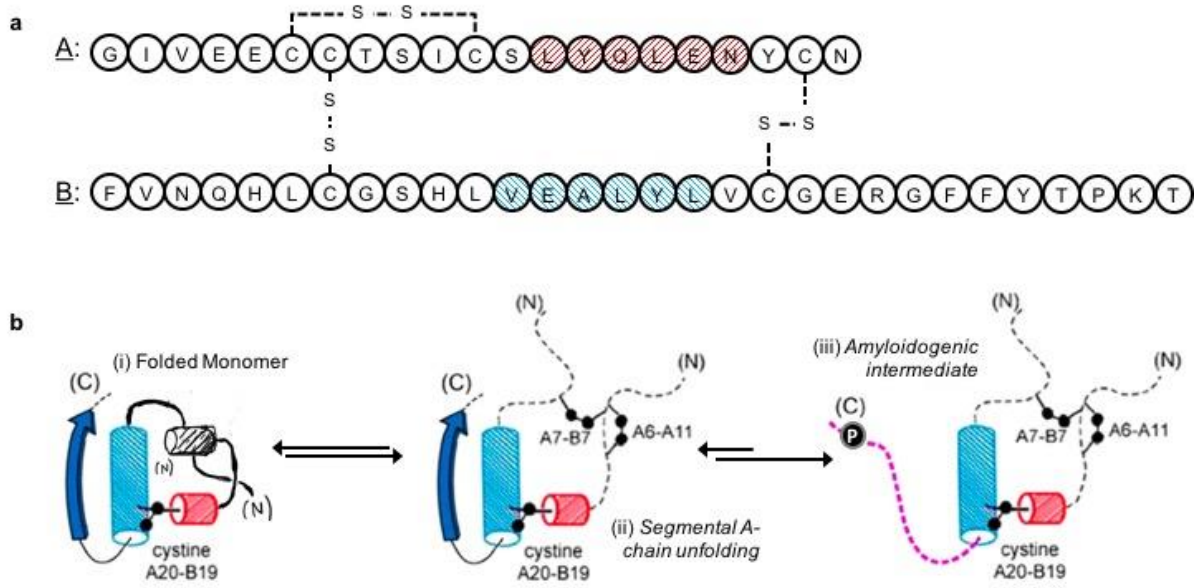
Hexameric FypI behaves like an RAI. Previous studies of RAIs have shown that destabilization of the dimer interface correlates with accelerated dissociation of the hexamer and rapid onset of insulin action[52,53]. Triggered dissociation of Zn^{2+}-hexamers by addition of the chelating agent terpyridine[54] (Figure 3.4c) revealed nearly identical rates of dissociation for FzpI and ProI, ($\tau_{1/2}$ = 98.3 ± 5.5 s and 90.4 ± 4.2 s, respectively), and for FypI and AspI ($\tau_{1/2}$ = 40.0 ± 4.1 s and 42.7 ± 4.3 s, respectively). dFpI exhibited hexamer dissociation kinetics ($\tau_{1/2}$ = 67.1 ± 4.3s) between those of ProI (and FzpI) and AspI (and FypI); the intermediate $\tau_{1/2}$ for dFpI suggests that the two fluoro-stereoisomers may have

opposing effects on hexamer dissociation rates. It should be noted that the rate of hexamer dissociation for the fluoroinsulins roughly correlates with K_D, aligning with previous studies performed on RAIs and in contrast with the results for the hydroxyinsulins (Chapter 2).

Fluorinated insulins fibrillate faster than (*4S*)-hydroxyinsulin. Given the abundance of literature detailing stabilizing effects due to fluorination in proteins[19,55,56], and the discovery of HzpI (Chapter 2) with enhanced resistance (~3-fold) against fibril formation, we speculated that fluorination at the proline could have similar results. Each of the fluoroinsulins was subjected to fibrillation lag time analysis (Figure 3.4d). We found that the orientation of the fluorine at position B28 did influence fibrillation time lag, but none of the effects were as striking as HzpI (Figure 3.4d; dotted grey line). In the case study for β2m, Fzp-β2m was found to be the most stable variant due to Fzp's $K_{trans/cis}$ of 2.5. When Pro32 is in its *cis* form, β2m is more resistant to fibril formation; however, this explanation does not suffice for insulin because the mechanism for insulin fibrillation does not depend on the *cis* orientation of ProB28 (Figure 3.6). The hypothesis for insulin fibrillation[57-62] is that two key segments[63] (^{13}LYQLEN^{18} of the A chain and ^{12}VEALYL^{17} of the B chain) when unfolded and exposed are responsible for the formation and growth of insulin fibrils (Figure 3.6). These two key segments are buried and unexposed when insulin is dimeric, leading to the hypothesis that the insulin dimer is more stable than the monomer. This assumption is further corroborated in the literature by experimental evidence that a covalently-linked, permanent insulin dimer is completely resistant to fibrillation[64].

Figure 3.6 | Fibril forming segments and hypothesis for insulin fibrillation.

***a**, Pictorial representation of mature insulin residues with fibril-forming segments colored in red (A chain) and blue (B chain). **b**, Hypothesis for insulin fibrillation. Schematic expanded from figure previously described[65].*

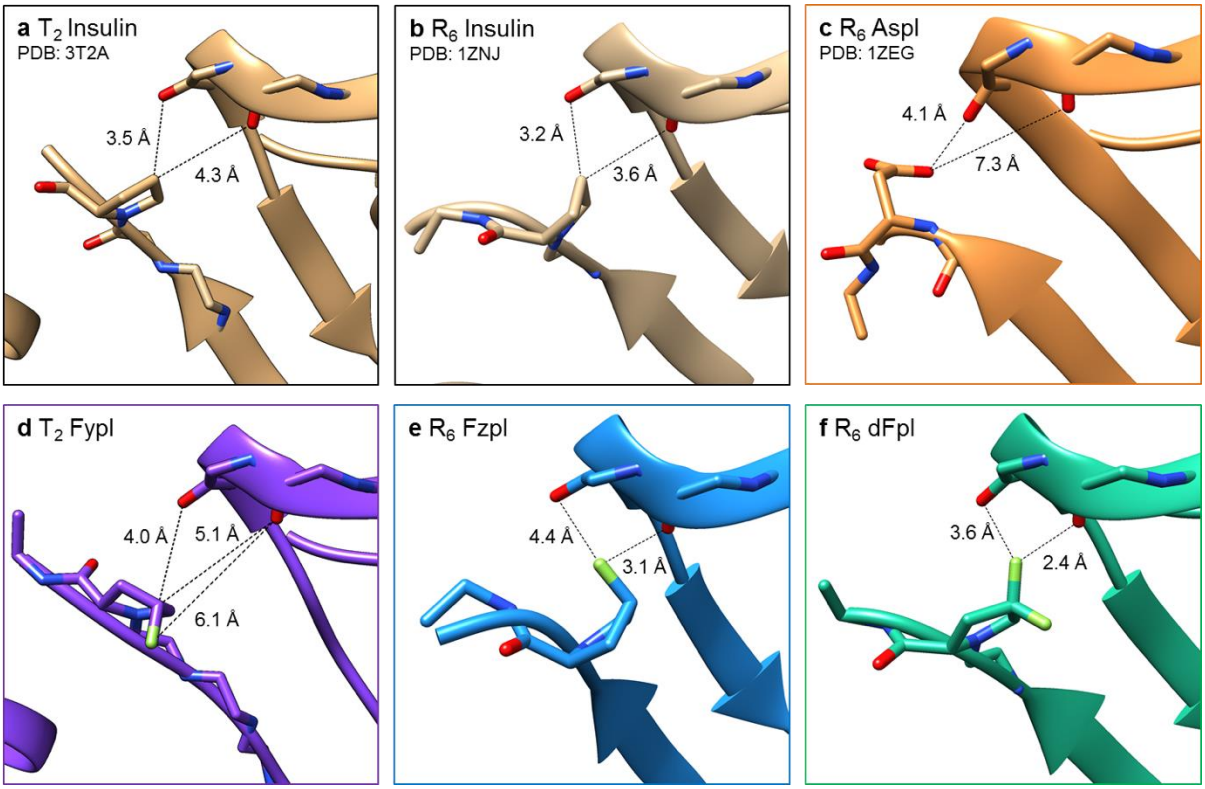


In Chapter 2, we speculated that the hydroxyl group of HzpB28 strengthens the dimer state through hydrogen bonding (Figure 2.8) and prevents the C-terminus of the B chain from fraying and hence, delaying fibrillation. Fluorine cannot form hydrogen bonds in the same manner as a hydroxyl group. Therefore, we suspect that the slight increase in fibrillation lag time for FzpI and dFpI, compared to ProI and AspI, may arise from the network of hydrophobic interactions[66] in the dimer interface (formed by the side chains of residues at positions B16, B24-25, B28 and others).

To help us further understand the atomistic interactions that are causing the biophysical behavior of the fluoroinsulins, we obtained crystal structures of all three variants (FzpI, FypI, and dFpI).

Figure 3.7 | Crystal structures of FzpI, FypI, and dFpI.

*In top panels (**a**, **b**, **c**), insulins from PDB in the T_2 dimer (**a**) and R_6 hexamer (**b**, **c**) forms. In bottom panels (**d**, **e**, **f**), structures of fluoro-insulins in T_2 dimer (**d**) and R_6 hexamer (**e**, **f**). All images, except (**c**), highlight the distance between the 4^{th}-carbon or its substitution group of the prolyl ring at position B28 and its closest neighbors, backbone carbonyl oxygen atoms of GlyB20' and GluB21. For AspI (**c**), the distances highlighted are between the backbone carbonyl oxygen atoms of GlyB20' and GluB21, and the closest oxygen atom from AspB28.*



Crystal structures reveal *endo*/*exo* conformations of the fluorinated prolyl ring. Solved crystal structures for fluoroinsulins, in comparison with ProI, display overall backbone root-mean-square-deviation (RMSD) values of 0.41 Å (R_6- FzpI), 0.39 Å (T_2- FypI), and 0.58 Å (R_6-dFpI)[67]. The overall crystal structures of R_6-FzpI and R_6-dFpI are similar to R_6-ProI; the prolyl rings at both FzpB28 (Figure 3.7e) and dFpB28 display (Figure 3.7f) the typical *endo* conformation seen at ProB28 (Figure 3.7ab). This structural similarity is expected for FzpI and ProI, due to their similar K_D values, hexamer dissociation kinetics, and fibrillation lag

times, but not for dFpI, whose biophysical characteristic values (Table 3.1) are intermediate between ProI (and FzpI), and AspI (and FypI).

Table 3.1 | Biophysical characteristics of fluoroinsulins.

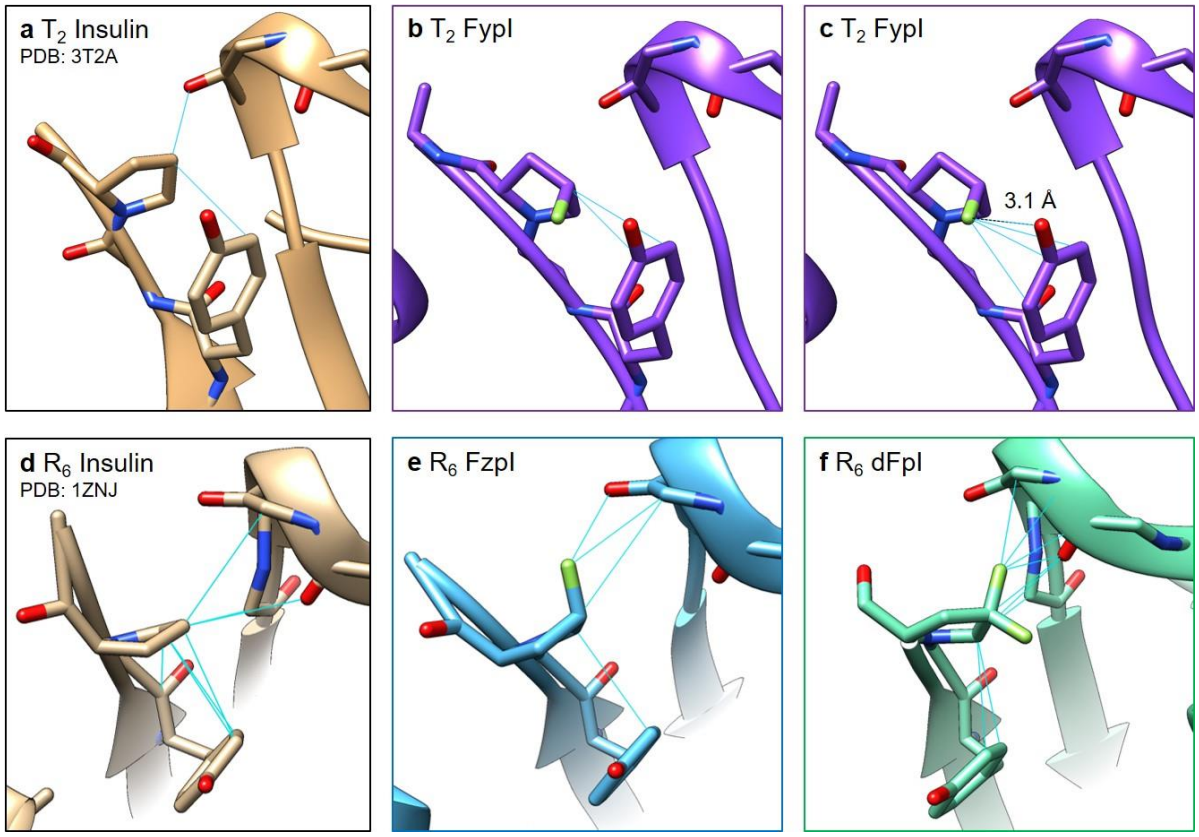
*Errors are given as one standard deviation (n ≥ 4). [*]Quantified by MALDI-MS on proinsulin peptide obtained by gluC digestion: [46]RGFFYT**P**KTRRE[57]. Samples were prepared from combining 5-1L shake flask expression and purification.*

B28 Amino Acid	Incorporation Level[*] (%)	Insulin Variant	Hexamer $\tau_{½}$ (s)	Fibrillation Lag Time (h)	K_D dimer (μM)
L-Proline	--	ProI	90.4 ± 4.2	5.1 ± 1.5	9 μM[68]
Fzp	97	FzpI	98.3 ± 5.5	8.4 ± 1.2	~2 μM
Fyp	96	FypI	40.0 ± 4.1	2.6 ± 0.3	~150 μM
dFp	86	dFpI	67.1 ± 4.3	9.0 ± 0.5	~150 μM
L-Aspartic acid	--	AspI	42.7 ± 4.3	5.3 ± 1.0	>500 μM[69]

One striking revelation of the solved fluoroinsulin structures is that, in the crystal structure of T_2-FypI (Figure 3.7d), FypB28's prolyl ring is in the *exo* conformation, which has yet to be seen in any crystal structures of insulin. The expected conformation for the prolyl ring at position B28 is the *endo* conformation (Figure 2.7, Figure 2.8 and Figure 3.7abef). An *exo* conformation of the prolyl ring results in a significant perturbation (from *endo*), causing both the γ-carbon and (*4R*)-fluorine of FypB28 to be further away from its nearest backbone neighbors (GluB21' and GlyB20') compared to T_2-ProI (Figure 3.7a), which may account for FypI's RAI-like behavior, and with similar distances measured in the structure of R_6-AspI[70] (Figure 3.7c). Although we were unable to obtain crystal structures of R_6-FypI to directly compare FypI to R_6-ProI and R_6-AspI, this *exo* conformation should also be present in the FypI hexamer, as the *endo* conformation for ProI is present in both its T_2 dimer and R_6 hexamer forms.

Figure 3.8 | Van der Waals contacts within dimer interface for fluoroinsulins.

*In top panels (**a**, **b**, **c**), images of T_2 insulins highlighting vdW contact partners for γ-carbon of the prolyl ring at position B28 (**a**, **b**) or fluorine at the 4th position (**c**). In bottom panels (**d**, **e**, **f**), images R_6 insulins highlighting vdW contact partners for carbons (α, β, γ, and δ) of the prolyl ring at position B28 (**d**, **e**, **f**) and fluorine at the 4th position (**e**, **f**). vdW contacts are represented by teal lines.*



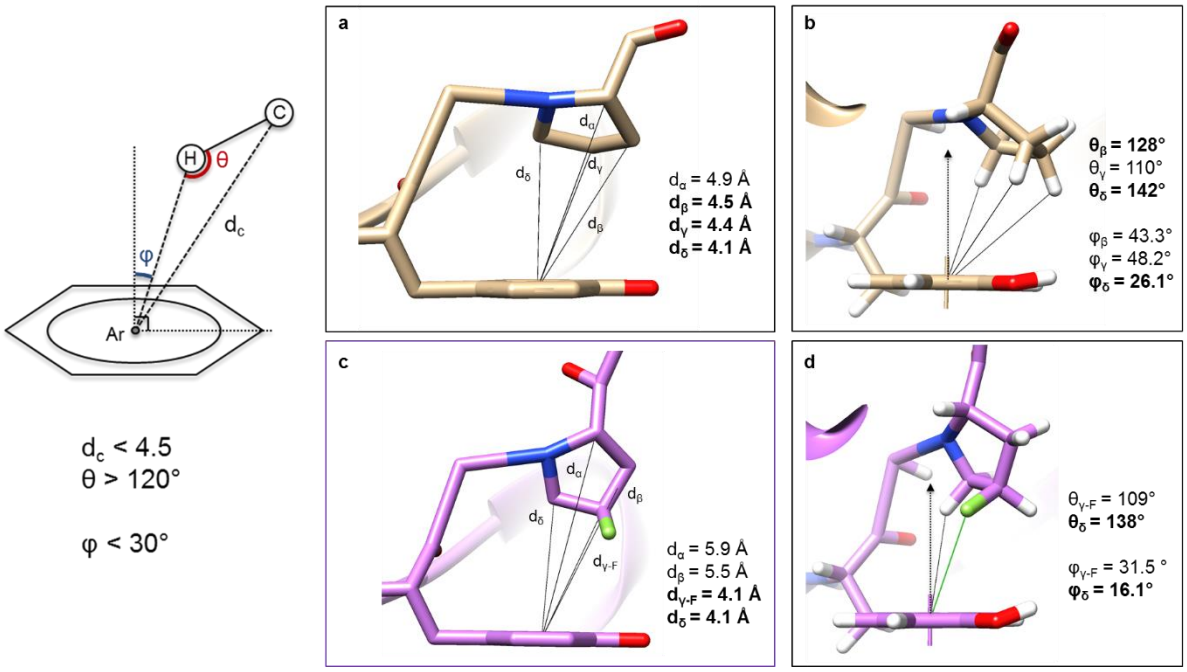
Crystal structures of fluorinated insulins reveal new van der Waals contacts. Hydrogen bonding interactions between backbone atoms from residues B24-B26 on the antiparallel β-strands of each insulin monomer[66] are the driving forces for dimerization of the insulin monomer. Dimerization is also aided by hydrophobic interactions between the β-turns (residues B20-23)[68] and C-terminal residues B27-29 of each insulin monomer[69]. We used crystallography software to determine the location of van der Waals (vdW) contacts[71] between the γ-carbon and fluorine group at position B28 and residues in the dimer interface. The prolyl ring of ProB28 in both the T_2 and R_6 structures form vdW contacts

primarily with residues GluB21′ and GlyB20′ of the interacting insulin monomer across the dimer interface (Figure 3.8ad, contacts for γ-carbon are shown). The vdW contacts for γ-carbon and (*4S*)-fluorine of FzpB28 in the R_6-FzpI structure are also primarily with GluB21′ and GlyB20′ (Figure 3.8e). The similar vdW contacts for R_6-ProI and R_6-FzpI are reflected in the biophysical properties, as FzpI has a dimerization constant similar to that of ProI and dissociates from its hexamer form at a rate indistinguishable from that of ProI. FzpI contains additional vdW contacts made by the (*4S*)-fluorine of FzpB28 to GluB21′ and GlyB20′ (Figure 3.8e), which may be reflected in the aggregation data (Figure 3.4d) as FzpI has a slightly delayed fibrillation lag time compared to ProI. Like R_6-FzpI, the γ-carbon and the (*4S*)-fluorine of dFpB28 contacts GluB21′ and GlyB20′ across the dimer interface, and the biophysical data of dFpI roughly correlates with these interactions, as its dimerization constant and hexamer dissociation rate fall between values for FzpI and ProI, and AspI, with a fibrillation lag time that indistinguishable from that of FzpI.

The (*4R*)-fluorine of dFpB28 also interacts with TyrB26 on the same B chain (Figure 3.8f); since this interaction is not present in R_6-FzpI, we theorize it may be responsible for the enhanced hexamer dissociation rate of dFpI compared to FzpI and ProI. Unfortunately, it is difficult to form any conclusions regarding the atomistic differences that may be responsible for the (lessened) propensity to fibrillate for FzpI and dFpI because we were unable to obtain crystal structures in the T form. However, we can speculate from the R_6 structures that the increase in stability for FzpI and dFpI may be due to the additional contacts across the dimer interface (with GluB21′ and GlyB20′) and with TyrB26 (Figure 3.8def), compared to R_6-ProI.

Figure 3.9 | CH-π interaction between ProB28 and TyrB26.

*Criteria for CH/π in the literature are assessed by: (d_C and ϑ)[72] or (d_C, ϑ, and ψ)[73,74]; the latter being more conservative but can exclude weaker CH/π interactions. Geometries of CH-π interactions between ∂-carbon of ProB28 and the centroid of the benzene ring of TyrB26 for in an ideal orientation (**left**), T_2-ProI (**a**, **b**), and T_2-FypI (**c**, **d**).*

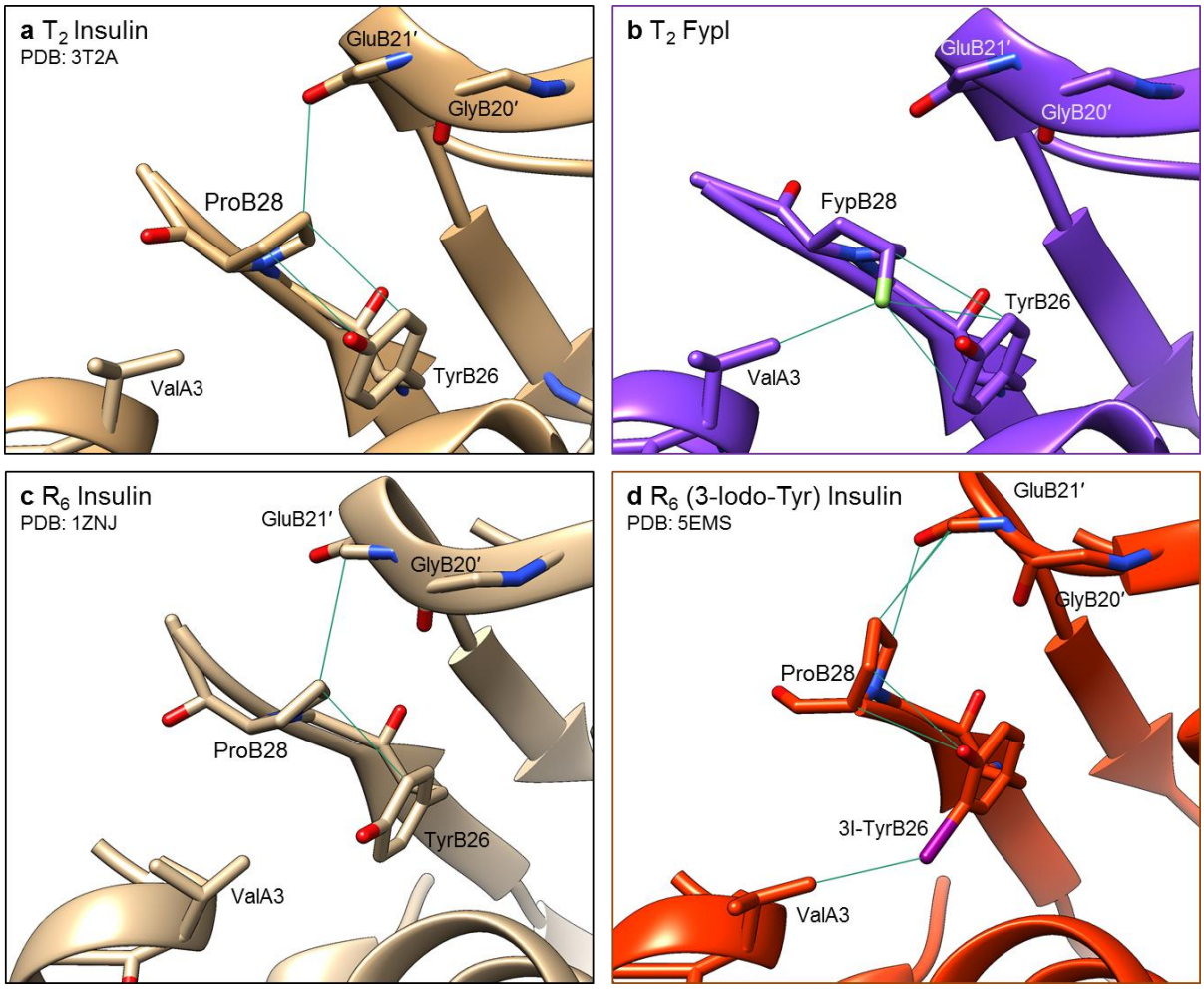


The (*4R*)-fluorine and γ-carbon at position B28 primarily interacts with the aromatic ring of TyrB26 in the T_2-FypI structure (Figure 3.8bc); FypB28 does not contact any residues across the dimer interface (residues B20' to B29'). Both computational[73] and experimental[75,76] evidence in the literature suggests the hydrogen from CH bonds of proline can interact favorably with the π-component of aromatic amino acids. The geometric orientation between (*4R*)-fluorine of FypB28 and the aromatic ring of TyrB26 is more indicative of a destabilizing interaction[77] because, not only is fluorine unable to form a CH/π interaction[78] but, the (*4R*)-fluorine of FypB28 may also replace a potential CH/π interaction present in the wild-type structure (Figure 3.9ab). From these observations, we hypothesize that the lack of vdW contacts in FypI's dimer interface, the presence[79] of (*4R*)-fluorine, and

absence of a CH/π interaction with TyrB26, through replacement of hydrogen at the γ-carbon of the B28 prolyl ring by (*4R*)-fluorine, may result in a destabilized dimer, faster hexamer dissociation kinetics, and an increased propensity to fibrillate.

Figure 3.10 | Crystal structures of halogenated insulins with vdW contacts with A chain.

*In top panels (**a**, **b**), T_2 insulins highlighting vdW contact partners for γ-carbon of the prolyl ring at position B28 (**a**, **b**) or fluorine at the 4th position (**b**). In bottom panels (**c, d**), R_6 insulins highlighting vdW contact partners for γ-carbon of the prolyl ring at position B28 (**c, d**) and iodine at the 3rd position of TyrB26 (**d**). vdW contacts are represented by solid teal lines.*



Halogen-based medicinal chemistry has been successful for pharmaceutical research and development, and there is much interest in extending this technology for engineering therapeutic proteins[80]. As protein engineers, corroborating information in the literature can

give significance to certain interactions that might seem minor in a single study. An example is the vdW contact between a halogen and residue ValA3; this interaction is present in both T_2-Fyp and the crystal structure of (3-Iodo-Tyr)-ProI in the R_6 hexamer state.

Through ncAA mutagenesis at ProB28 of insulin, we have obtained a greater understanding of the molecular interactions responsible for the tradeoffs exhibited between stability and hexamer dissociation rates for insulin. The information revealed by our fluorinated insulins provide further confirmation of the importance of hydrophobic interactions between residues along the dimer interface and its effects on biophysical properties such as dimerization, dissociation kinetics, and stability. We suggest that the placement of (*4R*)-fluorine of FypB28 near the aromatic ring of TyrB26 may be destabilizing and thereby, increases both the rate of disassembly and fibril formation. A facile strategy to perform atom-by-atom perturbations would be to use ncAA mutagenesis at either or both B26 or B28 positons to selectively strengthen CH/π interactions (e.g. through conversion to polar/π or cation/π interactions). For example, the use of a fluorinated aromatic amino acid[81,82] at position B26 would enhance the π-component of the aromatic ring or a more polar or positively-charged (*R*)-substituent at either or both the γ- or δ-carbon of positon B28 to obtain a stronger CH/π or cation-π interaction[83]. The use of ncAA mutagenesis in proteins is akin to the use of medicinal chemistry on small molecules, which has been the driving force for pharmaceutical research and development in the past century. Perhaps, as the push for more biologics to be available on the market increases, ncAA mutagenesis will prove to be a useful tool for performing medicinal chemistry on proteins.

Materials and Methods

Materials. All canonical amino acids and protected *Boc*-(4,4)-difluoro-L-proline were purchased from Sigma. *(4S)*-fluoro-L-proline (Hzp) and *(4R)*-fluoro-L-proline (Fyp) was purchased from Bachem Americas. All solutions and buffers were made using double-distilled water (ddH_2O).

Strains and plasmids. The proinsulin (PI) gene with an *N*-terminal hexa-histidine tag (6xHIS), and flanked by *Eco*R1 and *Bam*H1 cut sites was ordered as a gBlock (Integrated DNA Technologies). Both the gBlock and vector pQE80L for IPTG-inducible expression were digested with *Eco*RI and *Bam*HI. Linearized vector pQE80L was dephosphorylated by alkaline phosphatase (NEB). Ligation of the digested PI gene and linearized vector yielded plasmid pQE80PI (to produce ProI). To make plasmid pQE80PI-proS (to produce FzpI, FypI and dFpI): Genomic DNA was extracted from *E. coli* strain DH10β using DNeasy Blood and Tissue Kit (Qiagen). Primers (Integrated DNA Technologies) were designed to amplify the *E. coli proS* gene, encoding prolyl-tRNA synthetase, under constitutive control of its endogenous promoter, from purified genomic DNA, and to append *Nhe*I and *Nco*I sites. The digested *proS* gene was then inserted into pQE80PI between transcription termination sites by ligation at *Nhe*I and *Nco*I restriction sites. Proline-auxotrophic *E. coli* strain CAG18515 was obtained from the Coli Genetic Stock Center at Yale University. Prototrophic *E. coli* strain BL-21 was used for rich media expression of canonical insulins (ProI, AspI). Site-directed mutagenesis of pQE80PI at B28 was performed to make plasmid pQE80PI-asp, which differs from pQE80PI by three nucleotides that specify a single amino acid mutation to aspartic acid. All genes and plasmids were confirmed by DNA sequencing.

Protein expression. Plasmids pQE80PI and pQE80PI-asp were transformed into BL21 cells and grown on ampicillin-selective agar plates. A single colony was used to inoculate 5 mL of Luria-Bertani (LB) medium and grown overnight; the resulting saturated culture was used to inoculate another 1 L of LB medium. All expression experiments were conducted at 37°C, 200 RPM in shake flasks (Fernbach 2.8 L flasks, Pyrex®). Each culture was induced with 1 mM IPTG at mid-exponential phase (OD_{600} ~0.8). For incorporation of fluoroprolines (Fzp, Fyp and dFp), pQE80PI-proS was transformed into CAG18515 cells, which were grown on ampicillin-selective agar plates. To facilitate growth, a single colony was used to inoculate 25 mL of LB medium and the culture was grown overnight prior to dilution into 1 L of 1X M9, 20 aa medium (8.5 mM NaCl, 18.7 mM NH_4Cl, 22 mM KH_2PO_4, 47.8 mM Na_2HPO_4, 0.1 mM $CaCl_2$, 1 mM $MgSO_4$, 3 mg/L $FeSO_4$, 1 µg/L of trace metals (Cu^{2+}, Mn^{2+}, Zn^{2+}, MoO_4^{2-}), 35 mg/L thiamine hydrochloride, 10 mg/L biotin, 20 mM D-glucose, 200 mg/L ampicillin with 50 mg/L of L-amino acids, each). At an appropriate cell density (OD_{600} ~0.8), the culture was subjected to a medium shift; briefly, cells were centrifuged and washed with saline prior to resuspension into 0.8 L of 1.25X M9, 19 aa (1X M9, 20 aa medium without L-proline). After cells were further incubated for 30 min to deplete intracellular proline, 200 mL of 5X additives (1.5 M NaCl, 2.5 mM Fzp or Fzp, or 5mM dFp) was added to the culture. After another 15 min of incubation at 37°C to allow amino acid uptake prior to induction, IPTG was added to a final concentration of 1 mM. At the end of 2 h, cells were harvested by centrifugation and stored at -80°C until further use.

Cell lysis and refolding from inclusion bodies. Cells were thawed on the benchtop for 15 min prior to resuspension in lysis buffer (B-PER®, 0.5 mg/mL lysozyme, 50 U/mL benzonase

nuclease). Cells were gently agitated at RT for 1 h prior to centrifugation (10 000 g, 10 min, RT); supernatant was discarded and the pellet was washed thrice: once with wash buffer (2 M urea, 20 mM Tris, 1% Triton X-100, pH 8.0) and twice with sterile ddH_2O; centrifugation followed each wash and the supernatant was discarded. The final washed pellet containing inclusion bodies (IBs, ~50% PI) was re-suspended in Ni-NTA binding buffer (8 M urea, 300 mM NaCl, 50 mM NaH_2PO_4, pH 8.0) overnight at 4°C or at RT for 2 h, both with gentle agitation. The suspension was centrifuged to remove insoluble debris; the remaining pellet was discarded and the supernatant was mixed with pre-equilibrated Ni-NTA resin (Qiagen) at RT for 1 h in order to purify PI from the IB fraction. Unbound proteins in the IB fraction were collected in the flow-through (FT), and the resin was washed with Ni-NTA wash buffer (8 M urea, 20 mM Tris base, 5 mM imidazole, pH 8.0) and Ni-NTA rinse buffer (8 M urea, 20 mM Tris base, pH 8.0) prior to stripping PI from the resin with Ni-NTA elution buffer (8 M urea, 20 mM Tris base, pH 3.0). Fractions (IBs, FT, W, elution) were collected and run under reducing conditions on SDS-PAGE (Bis/Tris gels, Novex®); elution fractions containing PI were pooled and solution pH was adjusted to 9.6 with 6 N NaOH in preparation for oxidative sulfitolysis. Oxidative sulfitolysis was performed at RT for 4 h, with the addition of sodium sulfite and sodium tetrathionate (0.2 M Na_2SO_3, 0.02 M $Na_2S_4O_6$); the reaction was quenched by 10-fold dilution with ddH_2O. To isolate PI from the quenched solution, the pH was adjusted to between 3.5 and 4.5 by adding 6 N HCl dropwise; the solution became cloudy. The solution was centrifuged (10 000 g, 10 min, RT) and supernatant discarded. The PI pellet was then re-suspended in refolding buffer (0.3 M urea, 50 mM glycine, pH 10.6) and protein concentration was estimated by the bicinchoninic acid assay (BCA assay,

Pierce®). The concentration of PI was adjusted to 0.5 mg/mL. Refolding was initiated by addition of β-mercaptoethanol to a final concentration of 0.5 mM and allowed to proceed at 12°C overnight with gentle agitation (New Brunswick® shaker, 100 RPM). Post-refolding, soluble PI was harvested by adjusting the pH of the solution to 4-5 by dropwise addition of 6 N HCl and by high speed centrifugation to remove insoluble proteins. The supernatant was adjusted to pH 8-8.5 by dropwise addition of 6 N NaOH and dialyzed against fresh PI dialysis buffer (7.5 mM sodium phosphate buffer, pH 8.0) at 4°C with five buffer changes to remove urea. The retentate (PI in dialysis buffer) was then lyophilized and subsequently stored at -80°C until further processing. Typical yields were 25-50 mg PI per L of culture (25-30 mg/L for non-canonical PI, 40-50 mg/L for canonical PI expression in rich media)

Proteolysis and chromatographic (HPLC) purification. The dry PI powder was re-dissolved in water to a final concentration of 5 mg/mL PI (final concentration of sodium phosphate buffer is 100 mM, pH 8.0). Trypsin (Sigma-Aldrich) and carboxypeptidase-B (Worthington Biochemical) were added to final concentrations of 20 U/mL and 10 U/mL, respectively to initiate proteolytic cleavage. The PI/protease solution was incubated at 37°C for 2.5 h; proteolysis was quenched by addition of 0.1% trifluoroacetic acid (TFA) and dilute HCl to adjust the pH to 4. Matured insulin was purified by reversed phase high-performance liquid chromatography (HPLC) on a C_{18} column using a gradient mobile phase of 0.1% TFA in water (solvent, A) and 0.1% TFA in acetonitrile (ACN; solvent, B). Elution was carried from 0% B to 39% B with a gradient of 0.25% B per minute during peak elution. Fractions were collected and lyophilized, and the dry powder was re-suspended into 10 mM sodium phosphate, pH 8.0. Insulin-containing fractions were verified by matrix-assisted laser

desorption/ionization-mass spectrometry (MALDI-MS; Voyager MALDI-TOF, Applied Biosystems) and SDS-PAGE to ensure identify and purity. Typical yields were 5-10 mg insulin per 100 mg PI (FzpI and FypI ~10mg end, dFpI on ~5mg end). Fractions were stored at -80°C in 10 mM phosphate buffer, pH 8.0 until further use.

Verification of dFp, Fyp and Fzp incorporation levels and maturation. A 30 µL aliquot of PI solution (8 M urea, 20 mM Tris, pH 8) was subjected to cysteine reduction and alkylation (5 mM DTT, 55°C, 20 min; 15 mM iodoacetamide, RT, 15 min, dark) prior to 10-fold dilution into 100 mM NH_4HCO_3, pH 8.0 (100 µL final volume). Peptide digestion was initiated with 0.6 µL of gluC stock solution (reconstituted at 0.5 µg/µL with ddH_2O, Promega) at 37°C for 2.5 h. The reaction was quenched by adding 10 µL of 5% TFA and immediately subjected to C_{18} ZipTip (Millipore) peptide purification and desalting according to the manufacturer's protocol. Peptides were eluted in 50% ACN, 0.1% TFA; the eluent was then diluted three-fold into matrix solution (saturated α-cyanohydroxycinnamic acid in 50% ACN, 0.1% TFA) and analyzed by mass spectrometry (Voyager MALDI-TOF, Applied Biosystems). Incorporation levels were analyzed prior to and after refolding; incorporation percentage was calculated by comparing total AUC (area under the curve, arbitrary units) of the non-canonical peak (1575 Da for the proinsulin peptide containing FzpB28 or FypB28 or 1594 Da if dFpB28, ~5825 Da for intact FzpI and FypI or ~5840 Da for dFpI) with total AUC of its wild-type counterpart (1557 Da and 5808 Da, respectively). Maturation of FzpI, FypI, and dFpI was analyzed after HPLC purification. TFA (1.6 µL, 5%) was added to 15 µL mature insulin solution (10 mM phosphate buffer pH 8.0) and subjected to C_{18} ZipTip (Millipore) peptide purification and desalting per the manufacturer's protocol.

Reduction of blood glucose in diabetic animals. NODscid (NOD.CB17-*Prkdc*^{scid}/J) mice were obtained from Jax Mice (Bar Harbor, Maine). Mice were maintained under specific pathogen-free conditions, and experiments were conducted according to procedures approved by the Institutional Animal Care and Use Committee at the City of Hope. Adult (8- to 12-week-old) male NODscid mice were injected intraperitoneally (50 mg/kg/day for 3 consecutive days) with freshly prepared streptozotocin (STZ) in 0.05 M 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 (defined as >200 mg/dL), measured by using a glucomonitor (FreeStyle; Abbott Diabetes Care, Alameda, CA) in blood (10 µL) sampled from the lateral tail vein. Insulin analogs concentrations were determined from A_{280} measurements using a molar extinction coefficient of 6080 M^{-1} cm^{-1} and diluted to 100 µg/mL into a formulation buffer according to a previous report[84]. Insulin analogs in solution were injected subcutaneously at the scruff and blood glucose was measured at the indicated time points.

Circular Dichroism. Spectra were collected in a 1 cm quartz cuvette in 10 mM sodium phosphate buffer pH 8.0. Data were collected from 195 nm to 250 nm, with step size of 0.25 nm and averaging time of 1 s on a Model 410 Aviv Circular Dichroism Spectrophotometer; spectra were averaged over 3 repeat scans. A reference buffer spectrum was subtracted from the sample spectra for conversion to mean residue ellipticity. Insulin concentrations ranged from 3 µM to 200 µM.

Hexamer Dissociation Assay. Insulins were quantified by both UV absorbance (NanoDrop Lite, ThermoFisher) and BCA assay, and normalized to 125 µM insulin prior to dialysis against 50 mM Tris/perchlorate, 25 µM zinc sulfate, pH 8.0 overnight at 4°C using a D-tube

dialyzer (Millipore Corp.) with MWCO of 3.5 kDa. Aliquots of dialyzed insulin solution were mixed with phenol to yield samples of the following composition: 100 µM insulin, 20 µM zinc sulfate, 100 mM phenol. Dissociation was initiated by addition of terpyridine (Sigma-Aldrich) to a final concentration of 0.3 mM from a 0.75 mM stock solution. A Varioskan multimode plate reader (ThermoScientific) was used to monitor absorbance at 334 nm. Kinetic runs were done at least in triplicate, and the data were fit to a mono-exponential function using Origin software. Post assay insulin samples were pooled and sample quality was determined by SDS-PAGE.

Fibrillation Assay. Insulin samples (60 µM in 10 mM phosphate, pH 8.0) were centrifuged at 22 000 g for 1 h immediately after addition of thioflavin T (ThT) (EMD Millipore) to a final concentration of 1 µM. Samples were continuously shaken at 960 RPM on a Varioskan multimode plate reader at 37°C, and fluorescence readings were recorded every 15 min or 20 min for 48 h (excitation 444 nm, emission 485 nm). Assays were run in quadruplicate, in volumes of 200 µL in sealed (Perkin-Elmer), black, clear-bottom 96 well plates (Grenier BioOne).

Crystallographic Studies. Insulin crystals were obtained from sitting drop trays set using a Mosquito robot (TTP Labtech). Drops were set by mixing 0.4 µL insulin solution with 0.4 µL well solution. Cells were cryoprotected in a mother liquor containing 30% glycerol prior to looping and flash freezing in liquid nitrogen. Data were collected at SSRL beamline BL12-2 using a DECTRIS PILATUS 6M pixel detector. Initial indexing and scaling was performed with XDS; for some structures, data were re-scaled in alternative space groups using Aimless[85]. Initial phases were generated by molecular replacement in PHASER with 3T2A (5UOZ) or

1EV3 (5UQA and 5UU3)[86]. Structure refinement was carried out in Coot and Refmac5[87,88]. Data were deposited in the PDB with the following codes: 5UQA (R_6-FzpI), 5UOZ (T_2-FypI), 5UU3 (R_6-dFpI). All distances and contacts were computed using UCSF Chimera Crystallography Software.

References

1. Brange, J. et al. Monomeric insulins obtained by protein engineering and their medical implications. *Nature* **333**(6174): 679-682 (1988).
2. Ciszak, E. et al. Role of C-terminal B-chain residues in insulin assembly: the structure of hexameric LysB28ProB29-human insulin. *Structure* **3**(6): 615-22 (1995).
3. Holmgren, S.K., Bretscher, L.E., Taylor, K.M. & Raines, R.T. A hyperstable collagen mimic. *Chemistry & Biology* **6**(2): 63-70 (1999).
4. Shoulders, M.D., Kamer, K.J. & Raines, R.T. Origin of the stability conferred upon collagen by fluorination. *Bioorg. Med. Chem. Lett.* **19**(14): 3859-3862 (2009).
5. 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* **6**(5): e19425 (2011).
6. 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* **14**(9): 1053-1057 (2013).
7. Deepankumar, K., Nadarajan, S.P., Ayyadurai, N. & Yun, H. Enhancing the biophysical properties of mRFP1 through incorporation of fluoroproline. *Biochemical and Biophysical Research Communications* **440**(4): 509-514 (2013).
8. Müller, K., Faeh, C. & Diederich, F. Fluorine in pharmaceuticals: Looking beyond intuition. *Science* **317**(5846): 1881 (2007).
9. O'Hagan, D. Understanding organofluorine chemistry. An introduction to the C-F bond. *Chemical Society Reviews* **37**(2): 308-319 (2008).
10. Swallow, S. Chapter two - fluorine in medicinal chemistry. in *Progress in Medicinal Chemistry*, Vol. 54 (eds. Lawton, G. & Witty, D.R.) 65-133 (Elsevier, 2015).
11. Buer, B.C., de la Salud-Bea, R., Al Hashimi, H.M. & Marsh, E.N.G. Engineering protein stability and specificity using fluorous amino acids: The importance of packing effects. *Biochemistry* **48**(45): 10810-10817 (2009).
12. Odar, C., Winkler, M. & Wiltschi, B. Fluoro amino acids: A rarity in nature, yet a prospect for protein engineering. *Biotechnology Journal* **10**(3): 427-446 (2015).
13. Renner, C. et al. Fluoroprolines as Tools for Protein Design and Engineering. *Angew Chem Int Ed Engl* **40**(5): 923-925 (2001).
14. Biffinger, J.C., Kim, H.W. & DiMagno, S.G. The polar hydrophobicity of fluorinated compounds. *ChemBioChem* **5**(5): 622-627 (2004).

15. Zhou, P., Zou, J., Tian, F. & Shang, Z. Fluorine bonding — How does it work in protein−ligand interactions? *Journal of Chemical Information and Modeling* **49**(10): 2344-2355 (2009).
16. Hakoshima, T. Leucine Zippers. in *eLS* (John Wiley & Sons, Ltd, 2005).
17. Gunasekar, S.K. et al. N-Terminal Aliphatic Residues Dictate the Structure, Stability, Assembly, and Small Molecule Binding of the Coiled-Coil Region of Cartilage Oligomeric Matrix Protein. *Biochemistry* **48**(36): 8559-8567 (2009).
18. Tang, Y. et al. Stabilization of Coiled-Coil Peptide Domains by Introduction of Trifluoroleucine. *Biochemistry* **40**(9): 2790-2796 (2001).
19. Marsh, E.N.G. Fluorinated proteins: From design and synthesis to structure and stability. *Accounts of Chemical Research* **47**(10): 2878-2886 (2014).
20. Holzberger, B., Rubini, M., Möller, H.M. & Marx, A. A highly active DNA polymerase with a fluorous core. *Angew Chem Int Ed Engl* **49**(7): 1324-1327 (2010).
21. Panasik, N., Eberhardt, E.S., Edison, A.S., Powell, D.R. & Raines, R.T. Inductive effects on the structure of proline residues. *International Journal of Peptide and Protein Research* **44**(3): 262-269 (1994).
22. DeRider, M.L. et al. Collagen stability: Insights from NMR spectroscopic and hybrid density functional computational investigations of the efect of electronegative substituents on prolyl ring conformations. *JACS* **124**(11): 2497-2505 (2002).
23. Hodges, J.A. & Raines, R.T. Stereoelectronic effects on collagen stability: The dichotomy of 4-fluoroproline diastereomers. *JACS* **125**(31): 9262-9263 (2003).
24. Schmid, F.X. Prolyl isomerase: enzymatic catalysis of slow protein-folding reactions. *Annual Review of Biophysics and Biomolecular Structure* **22**(1): 123-143 (1993).
25. Lin, L.N. & Brandts, J.F. Isomerization of proline-93 during the unfolding and refolding of ribonuclease A. *Biochemistry* **22**(3): 559-563 (1983).
26. Roderer, D., Glockshuber, R. & Rubini, M. Acceleration of the rate-limiting step of Thioredoxin folding by replacement of its conserved cis-proline with (4S)-fluoroproline. *ChemBioChem* **16**(15): 2162-2166 (2015).
27. Lummis, S.C.R. et al. Cis-trans isomerization at a proline opens the pore of a neurotransmitter-gated ion channel. *Nature* **438**(7065): 248-252 (2005).
28. Higgins, K.A., Craik, D.J., Hall, J.G. & Andrews, P.R. Cis-trans isomerization of the proline residue in insulin studied by 13C NMR spectroscopy. *Drug design and delivery* **3**(2): 159-170 (1988).
29. Reimer, U. & Fischer, G. Local structural changes caused by peptidyl–prolyl cis/trans isomerization in the native state of proteins. *Biophys. Chem.* **96**(2–3): 203-212 (2002).
30. Lu, K.P., Finn, G., Lee, T.H. & Nicholson, L.K. Prolyl cis-trans isomerization as a molecular timer. *Nat Chem Biol* **3**(10): 619-629 (2007).
31. Fuller, D.R. et al. Cis→Trans Isomerization of Pro7 in Oxytocin Regulates Zn2+ Binding. *Journal of The American Society for Mass Spectrometry* **27**(8): 1376-1382 (2016).
32. Adams, E.J. & Luoma, A.M. The Adaptable Major Histocompatibility Complex (MHC) Fold: Structure and Function of Nonclassical and MHC Class I–Like Molecules. *Annual Review of Immunology* **31**(1): 529-561 (2013).

33. Smith, D.P., Ashcroft, A.E. & Radford, S.E. Hemodialysis-related amyloidosis. in *Protein Misfolding Diseases: Current and Emerging Principles and Therapies* (eds. Ramirez-Alvarado, M., Kelly, J.W. & Dobson, C.M.) (Wiley, 2010).
34. Eakin, C.M., Berman, A.J. & Miranker, A.D. A native to amyloidogenic transition regulated by a backbone trigger. *Nat Struct Mol Biol* **13**(3): 202-208 (2006).
35. Jahn, T.R., Parker, M.J., Homans, S.W. & Radford, S.E. Amyloid formation under physiological conditions proceeds via a native-like folding intermediate. *Nat Struct Mol Biol* **13**(3): 195-201 (2006).
36. Eichner, T. & Radford, S.E. Understanding the complex mechanisms of β(2)-microglobulin amyloid assembly. *The FEBS Journal* **278**(20): 3868-3883 (2011).
37. Barbet-Massin, E. et al. Fibrillar vs Crystalline Full-Length β-2-Microglobulin Studied by High-Resolution Solid-State NMR Spectroscopy. *JACS* **132**(16): 5556-5557 (2010).
38. Salwiczek, M., Nyakatura, E.K., Gerling, U.I.M., Ye, S. & Koksch, B. Fluorinated amino acids: compatibility with native protein structures and effects on protein-protein interactions. *Chemical Society Reviews* **41**(6): 2135-2171 (2012).
39. Torbeev, V.Y. & Hilvert, D. Both the cis-trans equilibrium and isomerization dynamics of a single proline amide modulate β2-microglobulin amyloid assembly. *Proc. Natl. Acad. Sci. U. S. A.* **110**(50): 20051-20056 (2013).
40. Torbeev, V., Ebert, M.-O., Dolenc, J. & Hilvert, D. Substitution of proline32 by α-methylproline preorganizes β2-microglobulin for oligomerization but not for aggregation into amyloids. *JACS* (2015).
41. Chiti, F. & Dobson, C.M. Protein misfolding, functional amyloid, and human disease. *Annu Rev. Biochem* **75**(1): 333-366 (2006).
42. Pandyarajan, V. et al. Biophysical optimization of a therapeutic protein by non-standard mutagenesis: studies of an iodo-insulin derivative. *J. Biol. Chem.* **289**: 23367-23381 (2014).
43. Vajo, Z. & Duckworth, W.C. Genetically engineered insulin analogs: diabetes in the new millenium. *Pharmacol Rev* **52**(1): 1-9 (2000).
44. Mirmira, R.G., Nakagawa, S.H. & Tager, H.S. Importance of the character and configuration of residues B24, B25, and B26 in insulin-receptor interactions. *J. Biol. Chem.* **266**(3): 1428-1436 (1991).
45. Steiner, T. et al. Synthetic biology of proteins: Tuning GFPs folding and stability with fluoroproline. *PLoS ONE* **3**(2): e1680 (2008).
46. Doi, M. et al. Characterization of collagen model peptides containing 4-fluoroproline; $(4(S)\text{-Fluoroproline-Pro-Gly})_{10}$ forms a triple helix, but $(4(R)\text{-Fluoroproline-Pro-Gly})_{10}$ does not. *JACS* **125**(33): 9922-9923 (2003).
47. Barth, D., Milbradt, A.G., Renner, C. & Moroder, L. A (4R)- or a (4S)-fluoroproline residue in position Xaa of the (Xaa-Yaa-Gly) collagen repeat severely affects triple-helix formation. *ChemBioChem* **5**(1): 79-86 (2004).
48. Holzberger, B. & Marx, A. Replacing 32 proline residues by a noncanonical amino acid results in a highly active DNA polymerase. *JACS* **132**(44): 15708-15713 (2010).
49. 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* **5**(7): 928-936 (2004).

50. Menting, J.G. et al. Protective hinge in insulin opens to enable its receptor engagement. *Proc. Natl. Acad. Sci. U. S. A.* **111**(33): E3395-404 (2014).
51. Pocker, Y. & Biswas, S.B. Self-association of insulin and the role of hydrophobic bonding: a thermodynamic model of insulin dimerization. *Biochemistry* **20**(15): 4354-4361 (1981).
52. Pandyarajan, V. & Weiss, M.A. Design of non-standard insulin analogs for the treatment of diabetes mellitus. *Curr. Diab. Rep.* **12**(6): 697-704 (2012).
53. Birnbaum, D.T., Kilcomons, M.A., DeFelippis, M.R. & Beals, J.M. Assembly and dissociation of human insulin and LysB28ProB29-insulin hexamers: a comparison study. *Pharm. Res.* **14**(1): 25-36 (1997).
54. Rahuel-Clermont, S., French, C.A., Kaarsholm, N.C. & Dunn, M.F. Mechanisms of stabilization of the insulin hexamer through allosteric ligand interactions. *Biochemistry* **36**(19): 5837-5845 (1997).
55. Kwon, O.-H. et al. Hydration dynamics at fluorinated protein surfaces. *Proc. Natl. Acad. Sci. U. S. A.* **107**(40): 17101-17106 (2010).
56. Yuvienco, C., More, H.T., Haghpanah, J.S., Tu, R.S. & Montclare, J.K. Modulating supramolecular assemblies and mechanical properties of engineered protein materials by fluorinated amino acids. *Biomacromolecules* **13**(8): 2273-8 (2012).
57. Brange, J. & Langkjoer, L. Insulin structure and stability. *Pharm Biotechnol* **5**: 315-50 (1993).
58. Brange, J., Andersen, L., Laursen, E.D., Meyn, G. & Rasmussen, E. Toward understanding insulin fibrillation. *J Pharm Sci* **86**(5): 517-525 (1997).
59. Brange, J., Dodson, G.G., Edwards, D.J., Holden, P.H. & Whittingham, J.L. A model of insulin fibrils derived from the x-ray crystal structure of a monomeric insulin (despentapeptide insulin). *Proteins: Structure, Function, and Genetics* **27**(4): 507-516 (1997).
60. Ahmad, A., Millett, I.S., Doniach, S., Uversky, V.N. & Fink, A.L. Partially folded intermediates in insulin fibrillation. *Biochemistry* **42**(39): 11404-11416 (2003).
61. Hua, Q.-X. & Weiss, M.A. Mechanism of insulin fibrillation: The structure of insulin under amyloidogenic conditions resembles a protein-folding intermediate. *J. Biol. Chem.* **279**(20): 21449-21460 (2004).
62. Ahmad, A., Uversky, V.N., Hong, D. & Fink, A.L. Early events in the fibrillation of monomeric insulin. *J. Biol. Chem.* **280**(52): 42669-42675 (2005).
63. Ivanova, M.I., Sievers, S.A., Sawaya, M.R., Wall, J.S. & Eisenberg, D. Molecular basis for insulin fibril assembly. *Proc. Natl. Acad. Sci. U. S. A.* **106**(45): 18990-18995 (2009).
64. Vinther, T.N. et al. Novel covalently linked insulin dimer engineered to investigate the function of insulin dimerization. *PLoS ONE* **7**(2): e30882 (2012).
65. Yang, Y. et al. An Achilles' heel in an amyloidogenic protein and its repair: Insulin fibrillation and therapeutic design. *J. Biol. Chem.* **285**(14): 10806-10821 (2010).
66. Baker, E.N. et al. The Structure of 2Zn Pig Insulin Crystals at 1.5 A Resolution. *Philos Trans R Soc Lond B Biol Sci.* **319**(1195): 369 (1988).
67. Marshall, H., Venkat, M., Seng, N.S., Cahn, J. & Juers, D.H. The use of trimethylamine N-oxide as a primary precipitating agent and related methylamine osmolytes as

cryoprotective agents for macromolecular crystallography. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **68**(Pt 1): 69-81 (2012).
68. Antolikova, E. et al. Non-equivalent role of inter- and intramolecular hydrogen bonds in the insulin dimer interface. *J. Biol. Chem.* **286**(42): 36968-77 (2011).
69. Brems, D.N. et al. Altering the association properties of insulin by amino acid replacement. *Protein Eng.* **5**(6): 527-533 (1992).
70. Whittingham, J.L., Edwards, D.J., Antson, A.A., Clarkson, J.M. & Dodson, G.G. Interactions of phenol and m-cresol in the insulin hexamer, and their effect on the association properties of B28 pro --> Asp insulin analogues. *Biochemistry* **37**(33): 11516-23 (1998).
71. Li, A.J. & Nussinov, R. A set of van der Waals and coulombic radii of protein atoms for molecular and solvent‐accessible surface calculation, packing evaluation, and docking. *Proteins: Structure, Function, and Genetics* **32**(1): 111-127 (1998).
72. Brandl, M., Weiss, M.S., Jabs, A., Sühnel, J. & Hilgenfeld, R. C-H···π-interactions in proteins. *J Mol Biol* **307**(1): 357-377 (2001).
73. Zondlo, N.J. Aromatic-Proline Interactions: Electronically Tunable CH/π Interactions. *Accounts of Chemical Research* **46**(4): 1039-1049 (2013).
74. Plevin, M.J., Bryce, D.L. & Boisbouvier, J. Direct detection of CH/π interactions in proteins. *Nat Chem* **2**(6): 466-471 (2010).
75. Holzberger, B., Obeid, S., Welte, W., Diederichs, K. & Marx, A. Structural insights into the potential of 4-fluoroproline to modulate biophysical properties of proteins. *Chemical Science* **3**(10): 2924-2931 (2012).
76. Lin, Y.-J., Chu, L.-K. & Horng, J.-C. Effects of the Terminal Aromatic Residues on Polyproline Conformation: Thermodynamic and Kinetic Studies. *The Journal of Physical Chemistry B* **119**(52): 15796-15806 (2015).
77. Bissantz, C., Kuhn, B. & Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *Journal of Medicinal Chemistry* **53**(14): 5061-5084 (2010).
78. Nakagawa, Y., Irie, K., Yanagita, R.C., Ohigashi, H. & Tsuda, K.-i. Indolactam-V Is Involved in the CH/π Interaction with Pro-11 of the PKCδ C1B Domain: Application for the Structural Optimization of the PKCδ Ligand. *JACS* **127**(16): 5746-5747 (2005).
79. Buer, B.C., Meagher, J.L., Stuckey, J.A. & Marsh, E.N.G. Structural basis for the enhanced stability of highly fluorinated proteins. *Proc. Natl. Acad. Sci. U. S. A.* **109**(13): 4810-4815 (2012).
80. El Hage, K. et al. Extending halogen-based medicinal chemistry to proteins: Iodo-insulin as a case study. *J. Biol. Chem.* **291**(53): 27023-27041 (2016).
81. Matsushima, A., Fujita, T., Nose, T. & Shimohigashi, Y. Edge-to-Face CH/π Interaction between ligand Phe-Phenyl and receptor aromatic group in the thrombin receptor activation. *J. Biochem.* **128**(2): 225-232 (2000).
82. Pace, C.J. & Gao, J. Exploring and Exploiting Polar–π Interactions with Fluorinated Aromatic Amino Acids. *Accounts of Chemical Research* **46**(4): 907-915 (2013).
83. Dougherty, D.A. Cation-π Interactions Involving Aromatic Amino Acids. *The Journal of Nutrition* **137**(6): 1504S-1508S (2007).

84. Pandyarajan, V. et al. Aromatic anchor at an invariant hormone-receptor interface: function of insulin residue B24 with application to protein design. *J. Biol. Chem.* **289**(50): 34709-27 (2014).
85. Winn, M.D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **67**(Pt 4): 235-242 (2011).
86. McCoy, A.J. et al. Phaser crystallographic software. *J Appl Crystallogr.* **40**(Pt 4): 658-674 (2007).
87. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **66**(Pt 4): 486-501 (2010).
88. Murshudov, G.N., Vagin, A.A. & Dodson, E.J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **53**(3): 240-255 (1997).

Acknowledgements

We thank J. T. Kaiser and P. Nikolovski of the Molecular Observatory at Caltech, and S. Russi and the scientific staff of Beamline 12-2 at the Stanford Synchrotron Radiation Laboratory for assistance. We thank S. Virgil of the Chemical Catalysis Center and M. Shahgholi of the Mass Spectrometry Facility at Caltech for their assistance. We thank T. Ku, J. Lebon, and J. Rawson at the City of Hope for performing insulin activity assays *in vivo*. We thank W. Glenn for editing this chapter.

This work and analysis was done in collaboration with Seth Lieblich. S.L. performed deprotection chemistry to obtain dFp; S.L. also performed experiments for insulin maturation and HPLC purification, sample preparation for circular dichroism spectra and in vivo mouse assays, and solving crystal structures of insulin.

Chapter 4 – Understanding the effects of changing ring composition at position B28

Abstract

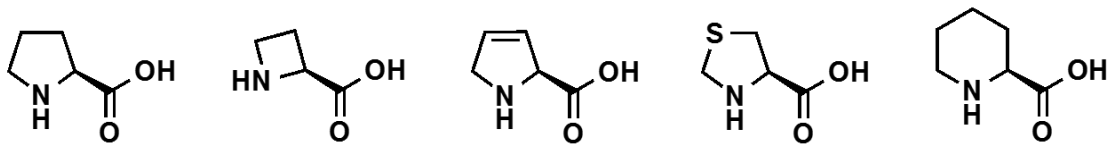
Insulin is known to be sensitive to mutation of the proline residue at position 28[1,2] of the insulin B-chain (ProB28). Rapid-acting insulins (RAIs) have taken advantage of this mutation to disrupt the inter-dimer interface and enhance the rate of disassembly from the insulin hexamer, but remain susceptible to fibrillation. Here we show that further replacement of ProB28 with proline analogs of changing ring composition results in minor perturbations structurally but can lead to large changes in insulin stability and disassembly. The changes in biophysical properties of insulin resulting from the removal, replacement or addition of carbon and hydrogen atoms at ProB28 demonstrates that it is possible to specifically modulate the biophysical properties of a therapeutic protein without adversely affecting biological activity.

Introduction

Prolyl ring. The uniqueness of proline primarily stems from the 5-membered prolyl ring, which is the result of the covalent bond between the nitrogen atom of the backbone and the α-carbon. Proline's restricted backbone limits accessible dihedral angles; thus, proline is found most often in β-turns and at the ends of α-helices[3]. Bioinformatic analysis have also revealed that proline residues are frequently found in repeating motifs that have diverse and critical functions biologically[4]. The restricted torsion angles and ring shape of the amino acid often introduce a "kink"[5] in the protein structure that can have dynamic and biologically relevant consequences. For example, it has been demonstrated in the literature that proline acts as a switch in transmembrane receptors[6] for initiating signaling pathways.

Figure 4.1 | Proline analogs for study of altered ring composition.

Compound **1***: L-proline; Compound* **2***: (*S*)-azetidine-2-carboxylic acid (Aze); Compound* **3***: (3,4)-dehydro-L-proline (Dhp); Compound* **4***: (1,3)-thiazolidine-4-carboxylic acid (Thz); Compound* **5***: (S)-piperidine-2-carboxylic acid* (Pip).



Protein engineering with proline analogs of varied ring composition. A significant body of work examines proline analogs, a small subset[7-11] of which have explored modifications to the prolyl ring. Initial studies on proline analogs of altered ring composition involved understanding the mechanisms governing their *cis*-trans isomerization preferences[12,13]; later, studies were expanded to the replacement of proline residues in proteins to understand the effects of *cis-trans* isomerization on protein properties.

Collagen. In addition to research on hydroxy- and fluoro-prolines, the presence of (*S*)-azetidine-2-carboxylic acid (Aze) in place of proline in collagen has also been studied extensively[14,15]. Evolutionary theories[16] have postulated that Pro was adapted as one of the twenty canonical amino acids (cAA) instead of the 4-membered Aze or the 6-membered (*S*)-piperidine-2-carboxylic acid (Pip), due to the more favorable backbone geometry resulting from a 5-membered ring. Consistent with such theories, replacing Pro with Aze in collagen led to detrimental stability effects and decreased activity[15].

Annexin V. One of the first detailed studies involving (1,3)-thiazolidine-4-carboxylic acid (Thz)[11] was done in the human protein annexin V, an anticoagulant[17], through residue-specific replacement of Pro. NMR spectroscopy of a Thz-containing peptide determined that Thz prefers the *cis*-amide ($K_{trans/cis}$ of 2.8) peptide bond[18] about 2-fold more than L-proline (Pro). Given that Thz (with similar bond lengths and angles to Pro) is considered to be isosteric to Pro, it is not surprising that the secondary structure and thermodynamic properties of Thz-annexin V are roughly unchanged compared to its wild-type protein[11].

Proteins with a C-terminal SsrA tag. The "unpuckered" proline analog (3,4)-dehydro-L-proline (Dhp)[19], Aze, and Thz were all incorporated into the model protein YbeL to assess the effects of proline analogs on protein degradation through SsrA tagging[10]. The exact function of YbeL is unknown. SsrA appends a peptide sequence (called the SsrA tag, which is recognized by multiple endogenous proteases) at the C-terminus to mark a protein for degradation[20]. The last four residues (prior to the stop codon) for YbeL contain two Pro residues. It was found that SsrA tagging is sensitive to the identity of the second last Pro residue in YbeL, where the tagging rate decreases in the order of: Dhp, Pro, Thz, and Aze;

however, the authors could not rationalize the underlying mechanisms for the tagging by SsrA with the composition of the proline analog-containing YbeL[10].

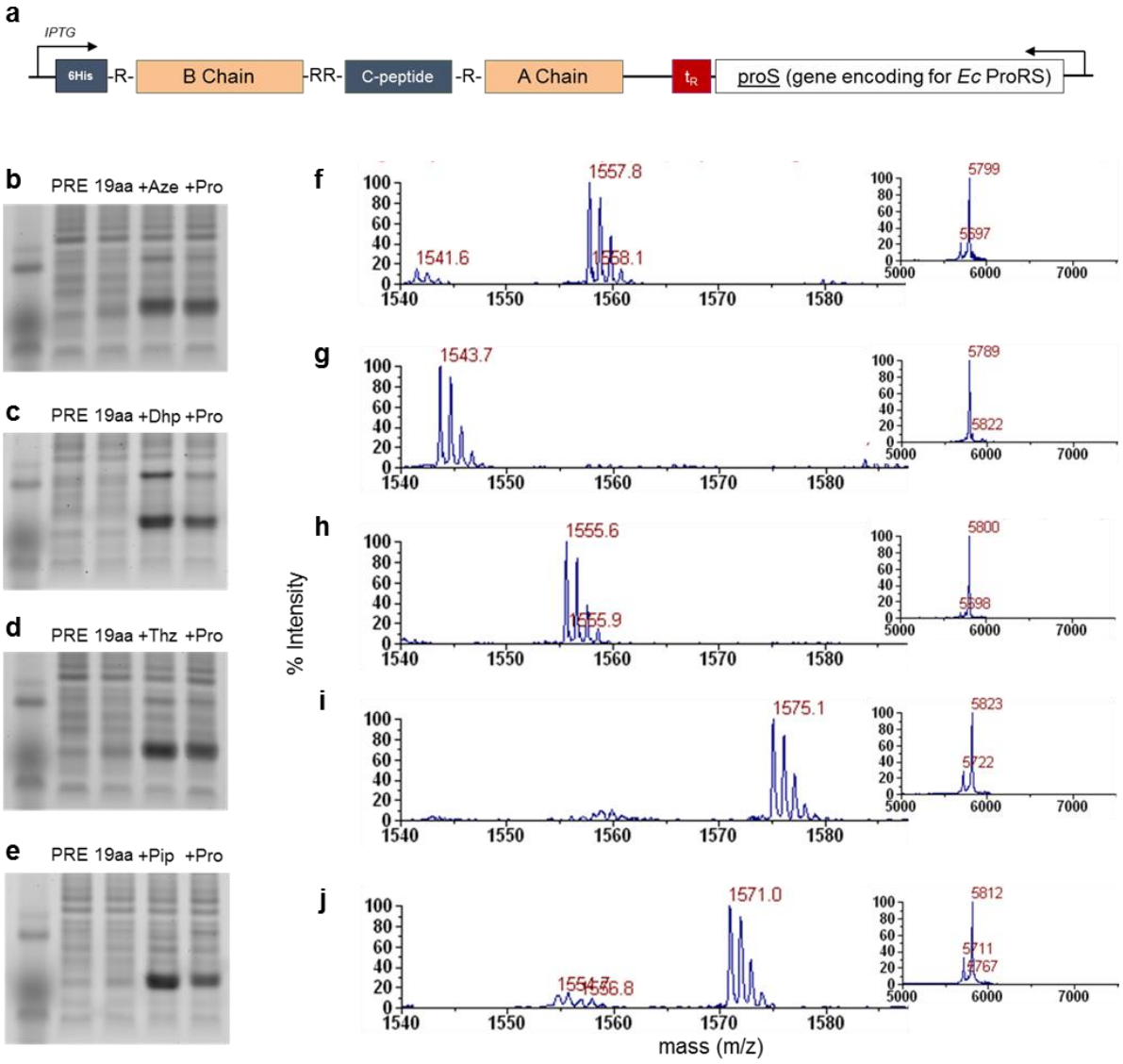
The analogs shown in Figure 4.1 are nonpolar and cannot form hydrogen bonds, and have been previously used as probes for understanding the role of proline residues in proteins. Thus, we propose to incorporate proline analogs with different prolyl ring sizes to modulate the biophysical properties of insulin and further investigate the nonpolar interactions present in the insulin dimer interface.

Results and Discussion

Recombinant production of AzeI, DhpI, ThzI and PipI. It has been reported[21] that Aze, Dhp, and Thz can be incorporated into newly synthesized proteins in *E. coli* using an overexpressed prolyl-tRNA synthetase (ProRS) under osmotic stress conditions, and Pip can be incorporated using a mutant ProRS (C443G). Incorporations were done using proline-auxotrophic *E. coli* strain CAG18515 for Aze and Pip, and *KS32* for Dhp, Thz. The *KS32* strain is the proline-auxotrophic version of parent strain JT31[22], which is deficient for proline dehydrogenase, that cannot metabolize Dhp and Thz[23] or synthesize Pro endogenously. Expression of PI was confirmed via polyacrylamide gel electrophoresis (Figure 4.2b-f). Incorporation levels (Table 4.1) were assessed using matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS; Figure 4.2f-j) and found to be approximately 90% by comparing the areas of the peaks corresponding to the non-canonical peptide relative to the wild-type peptide. Similarly, post-HPLC fractions containing mature insulin were verified by whole protein MALDI-MS (Figure 4.2 f-j, insets). PI and insulin yields were similar to the hydroxy- and fluoroinsulins described in Chapters 2-3.

Figure 4.2 | Insulin expression and incorporation of prolines analogs of changing ring size.

a, *Plasmid pQE80PI-proS containing gene construct proinsulin under inducible lac promoter and an overexpressed Ec ProRS for bacterial expression.* ***b-e,*** *SDS-PAGE of cell lysates with lanes labeled for pre-induction (PRE) and post-induction in minimal media supplemented with either nothing (19aa), Aze (**b**), Dhp (**c**), Thz (**d**) or Pip (**e**), or Pro at 0.5 mM.* ***f-j:*** *MALDI-MS traces of isolated proinsulin peptide fragment* ^{46}*RGFFYT**P**KTRRE*^{57} *obtained by gluC digestion. Peptide fragment masses corresponds to either wild type mass 1558 Da (**f**) or shifted masses if ncPro is incorporated (**g**, 1544 for Aze; **h**, 1555 Da for Dhp; **i**, 1575 Da for Thz; **j**, 1571 for Pip). Inset is whole protein MALDI-MS with observed masses: 5799 Da (**f**, ProI), 5789 Da (**g**, AzeI), 5800 Da (**h**, DhpI), 5823 Da (**i**, ThzI) and 5812 Da (**j**, PipI). All intact MALDI-MS spectra give experimental masses within error range (~10 Da or 1-2%) of its calculated value. All MALDI-MS spectra contain ion counts >10^3*



Changing ring compositions at B28 does not result in a destabilized insulin dimer. In the absence of Zn^{2+} and phenolic preservatives, ProI dimerizes with a dissociation constant (K_D) of approximately 10 µM. Monomeric forms of insulin give rise to characteristic circular dichroism (CD) spectra with distinct minima at 208 and 222 nm (e.g., AspI; Figure 4.3a). Dimerization causes a loss of negative ellipticity at 208 nm (e.g., ProI; Figure 4.3a). At 60 µM, all of the ring-size variants of insulin appear to be dimeric (with CD spectra similar to ProI, Figure 4.3a).

Table 4.1 | Characteristics of insulin variants with different prolyl ring compositions.

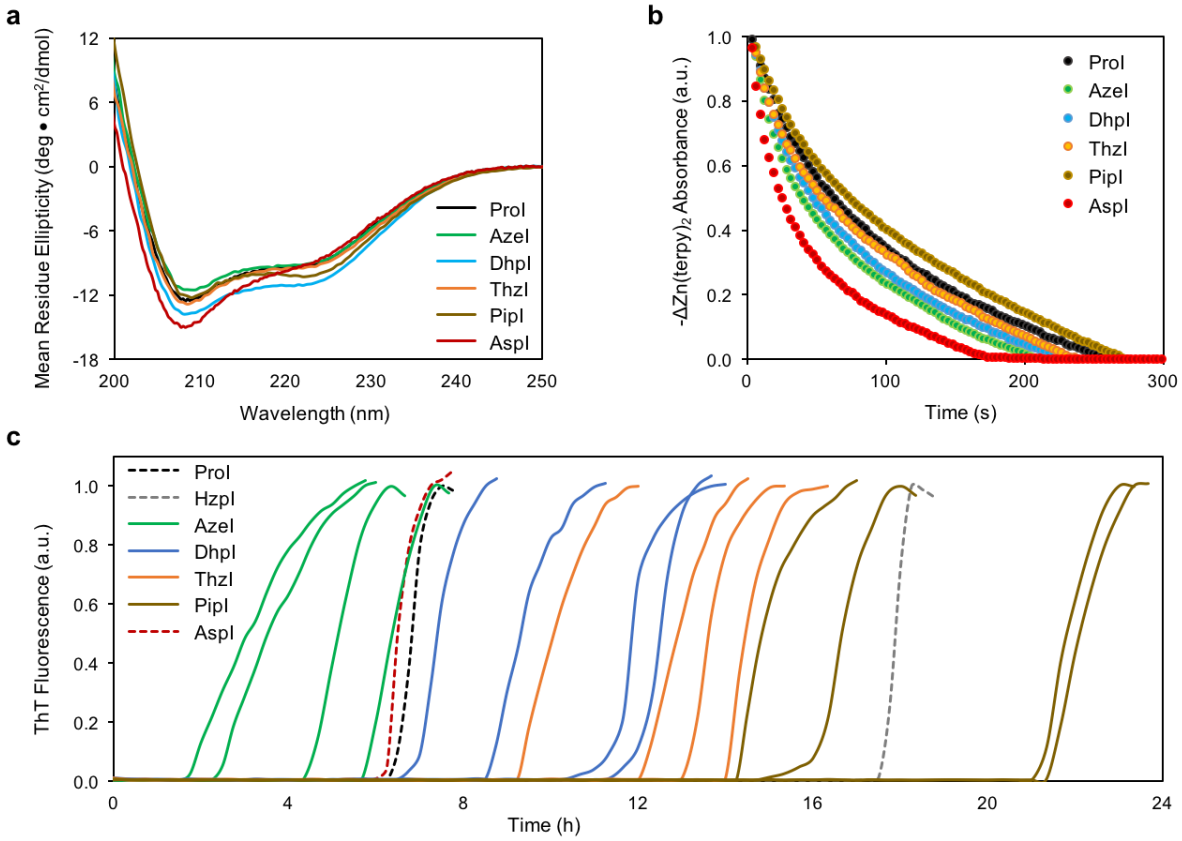
*Errors are given as one standard deviation (n ≥ 4). *Quantified by MALDI-MS on proinsulin peptide obtained by gluC digestion: ^{46}RGFFYT**P**KTRRE^{57}. Samples were prepared from combining 5-1L shake flask expression and purification.*

B28 Amino Acid	Incorporation Level* (%)	Insulin Variant	Hexamer $\tau_{½}$ (s)	Fibrillation Lag Time (h)	K_D dimer (µM)
L-Proline	--	ProI	90.4 ± 4.2	5.1 ± 1.5	9 µM[24]
AzeI	≤100%	AzeI	64.5 ± 4.5	4.5 ± 1.5	~10 µM
DhpI	≤100%	DhpI	72.3 ± 1.3	10.7 ± 1.5	~10 µM
ThzI	>85%	ThzI	83.9 ± 1.3	12.8 ± 1.9	~10 µM
PipI	93%	PipI	117 ± 19	18.8 ± 3.6	<2 µM
L-Aspartic acid	--	AspI	42.7 ± 4.3	5.3 ± 1.0	>500 µM[25]

To obtain estimates for the insulin dimer's dissociation constant (K_D)[26], we obtained CD spectra at varying concentrations. AzeI, DhpI and ThzI all have a K_D of approximately 10 µM, which is indistinguishable from that of ProI, indicating that slight perturbations to or removal of one carbon from the prolyl ring does not affect insulin dimerization. PipI, on the other hand, is more dimeric than ProI: at an insulin concentration of 2 µM, ProI is mostly monomeric while PipI is still primarily dimeric (K_D << 2 µM).

Figure 4.3 | Changing ring composition at position B28 affects dimerization and hexamer dissociation.

__a,__ Far UV CD spectra collected on 60 µM insulins in 10 mM phosphate buffer, pH 8.0 at 25°C. __b,__ Insulin hexamer dissociation following sequestration of Zn^{2+} by terpyridine. Zn^{2+}-(terpy) signal was monitored at 334 nm and fitted to a mono-exponential decay. __c,__ Representative fibrillation curves for 60 µM insulins (37°C, 960 RPM; n=4). Insulin fibrils were detected by the rise in Thioflavin T (ThT) fluorescence that occurs upon binding to fibrillar aggregates.



Decreasing size of prolyl ring at B28 speeds up hexamer dissociation and onset of fibrillation.

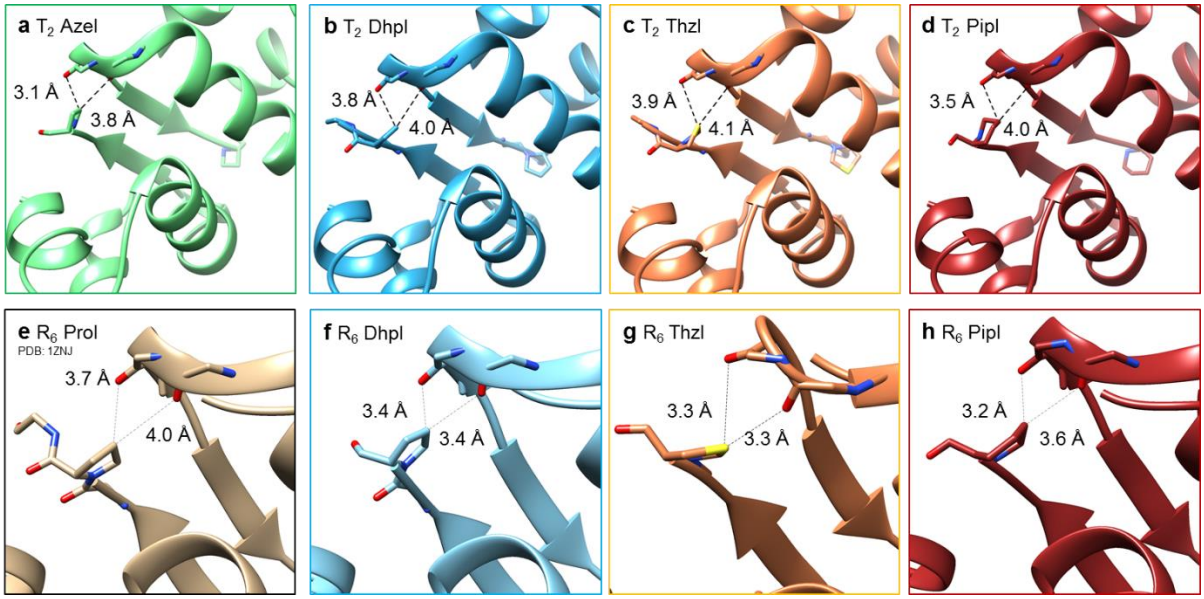
The increasing rate of hexamer dissociation roughly correlates with the decreasing size of the prolyl ring. AzeI, which has a 4-membered ring at AzeB28, dissociates the fastest (with a $\tau_{½}$ of 64.5 ± 4.5 s), followed by DhpI (with a $\tau_{½}$ of 72.3 ± 1.3 s), then ThzI (with a $\tau_{½}$ of 83.4 ± 1.3 s), and lastly PipI, which has a 6-membered ring at PipB28, dissociates the slowest (with a $\tau_{½}$ of 117 ± 19 s). Both AzeI and DhpI display faster hexamer dissociation rates than than ProI (with a $\tau_{½}$ of 90.4 ± 4.2 s), and have dimerization constants indistinguishable from that of ProI,

further demonstrating that it is possible to decouple the enhancement of hexamer dissociation rates ($\tau_{½}$) from destabilization of the insulin dimer (K_D).

Each of the insulin variants were subjected to fibrillation lag time analysis at dilute concentrations with the addition of amyloid stain Thioflavin T (ThT). We found that increasing size of the prolyl ring size at position B28 of insulin roughly corresponded to a slower fibrillation lag time (Figure 4.3c). Compared to ProI and AspI, AzeI is less stable, while DhpI, ThzI and PipI are more resistant to fibril formation. To help us understand the mechanisms of hexamer dissociation and fibril formation, we solved the crystal structures of AzeI, DhpI, ThzI, and PipI in the T_2 state and DhpI, ThzI, and PipI in the R_6 state.

Figure 4.4 | Crystal structures of AzeI, DhpI, ThzI, and PipI.

*In top panels (**a**, **b**, **c**, **d**), insulin variants in the T_2 dimer forms. In bottom panels (**e**, **f**, **g**, **h**), structures of insulins in R_6 hexamer. All images highlight the distance between the γ-carbon or its substitution group of the prolyl ring at position B28 and its closest neighbors, backbone carbonyl oxygen atoms of GlyB20' and GluB21', across the dimer interface.*

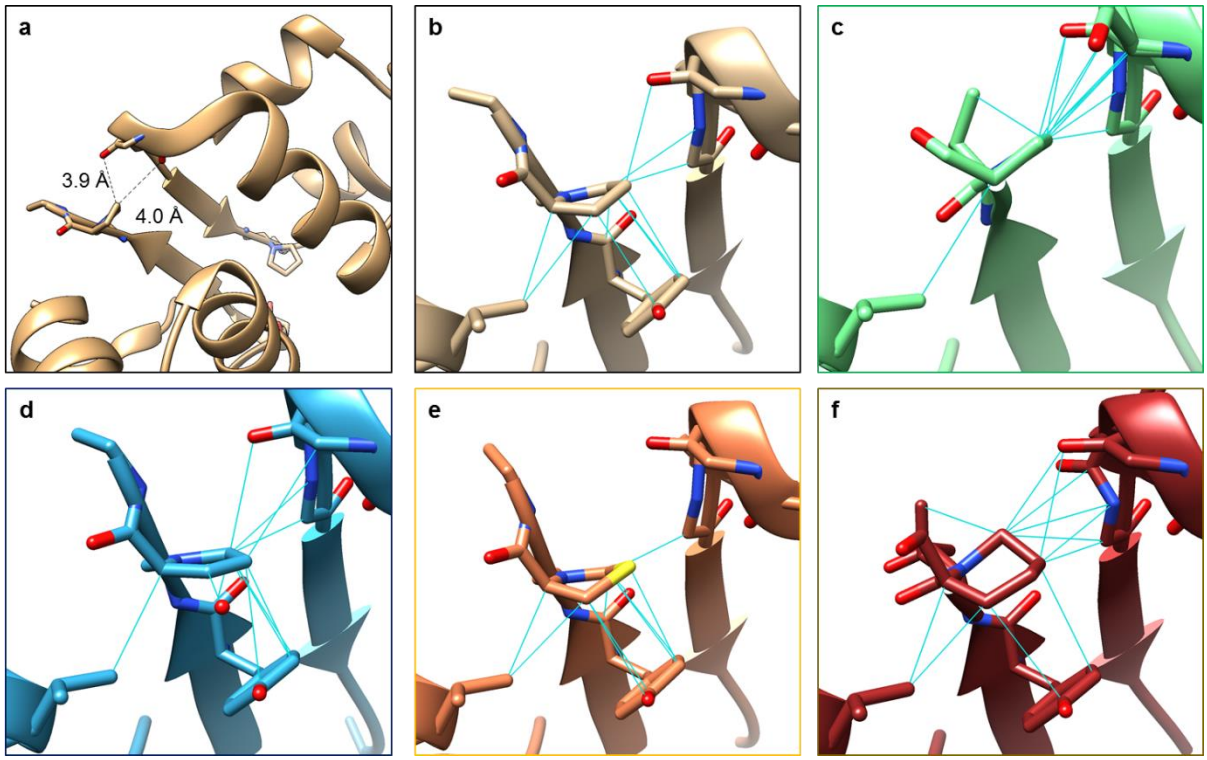


Crystal structures reveal minimal perturbations from changing prolyl ring size. Solved crystal structures for ring-size insulin variants, in comparison with ProI, display overall

backbone root-mean-square-deviation (RMSD) values of 0.41 Å (T_2-AzeI), 0.34 Å and 0.38 Å (T_2- and R_6-DhpI), 0.31 Å and 0.53 Å (T_2- and R_6-ThzI), and 0.28 Å and 0.27 Å (T_2- and R_6-PipI)[27]. From the crystal structures (Figure 4.4 and Figure 4.5a), we found that changing the size of the prolyl ring did not result in significant perturbations to the proline backbone trajectory and largely retains the overall structure of insulin. From the bond angles and molecular constraints[16,19,28] of Aze, Dhp, Thz and Pip, we found, unsurprisingly, that AzeB28 and DhpB28 are planar, ThzB28 has an *endo* conformation, and PipB28 prefers the chair conformation.

Figure 4.5 | Van der Waals contacts at position B28 with neighboring atoms in the T-state.

a, ***b***, *Crystal structures of T_2-ProI. Upper left image (**a**) highlights the distance between the γ-carbon or its substitution group of the prolyl ring at position B28 and its closest neighbors, backbone carbonyl oxygen atoms of GlyB20' and GluB21', across the dimer interface.* ***b-f***, *Images of T_2 insulins highlighting vdW contact partners for carbons of the prolyl ring at position B28 for ProI (**b**), AzeI (**c**), DhpI (**d**), ThzI (**e**), and PipI (**f**). vdW contacts are represented by teal lines.*

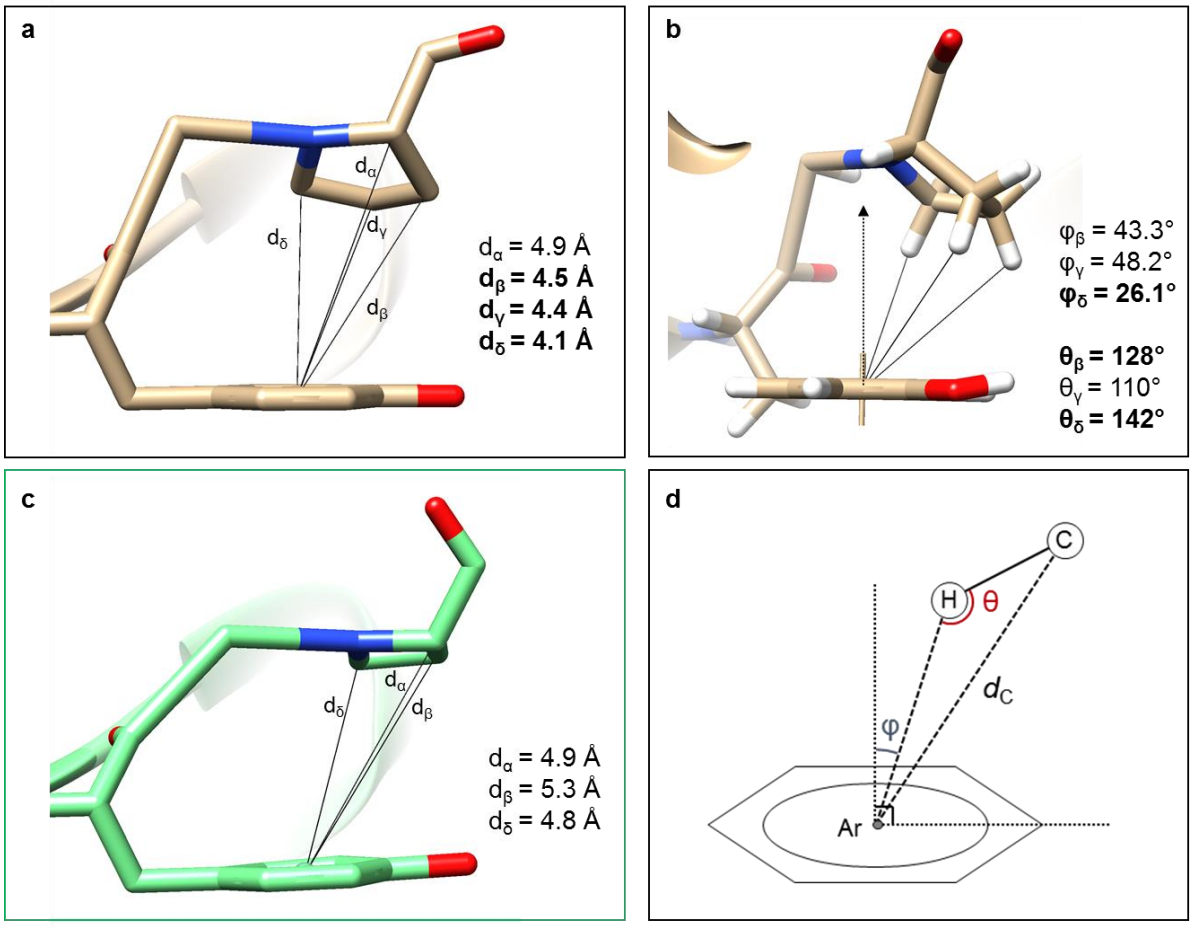


In Chapter 3, we investigated van der Waals (vdW) contacts between residues in the dimer interface and interactions with TyrB26, and found that minor perturbations could lead to significant biophysical consequences. Based on the biophysical data for hexamer dissociation and insulin fibrillation, we hypothesized that increasing ring size, as with Pip, provides more accessible surface area across the dimer interface, and decreasing ring size, as with Aze and Dhp, provides less accessible surface area for vdW contacts to occur. We used crystallography software[29] to calculate possible van der Waals contacts at each carbon at position B28: α-, β-, and δ-carbons for AzeI, α-, β-, γ-, and δ-carbons for DhpI and ThzI, and α-, β-, γ-, δ-, and ε-carbons for PipI (Figure 4.5b-f). PipI, with a significantly slower hexamer dissociation rate compared to ProI, and the most resistant to fibril formation (of the four ring-size variants), has an extensive network of vdW contacts in the dimer interface, and with neighboring residues TyrB26 and ValA3. Compared to PipI, DhpI and ThzI have more contacts with neighboring residues TyrB26 and ValA3 than interactions with residues GlyB20' and GluB21'. The network of vdW contacts for ProI, DhpI and ThzI in the T_2 state are qualitatively similar; however, it is difficult to quantify[30] the strength of hydrophobic interactions in proteins and thus, hard to rationalize the minor atomistic differences that account for the changes in biophysical behavior between ProI, DhpI, and ThzI. The structure of T_2-AzeI is unique (compared to the other ring-size insulin variants) because AzeB28 does not form any contacts with TyrB26 (Figure 4.5c). Structural and geometric analysis of the interactions between ProB28 and TyrB26 suggests that there exists a stabilizing CH/π interaction[31-34] between the δ-carbon, and likely also the γ-carbon, of ProB28 and the aromatic ring of TyrB26 (Figure 4.6ab), respectively. Aze's ring is shifted

further away from TyrB26, such that the distances, between the hydrogens of AzeB28 and the π-component of TyrB26's aromatic ring, are too large for any vdW overlap and thus, no stabilizing CH/π interactions are present in AzeI (Figure 4.6c). The absence of a CH/π interaction was also observed in the T_2-FypI structure; interestingly, FypI is the only other insulin variant we've produced that is less stable than ProI based on the results of the ThT fibrillation assay. The T_2 structures of DhpI, ThzI and PipI shown that CH/π interactions with TyrB26 were retained (analyses not shown).

Figure 4.6 | AzeB28 does not retain any CH-π interactions with TyrB26.

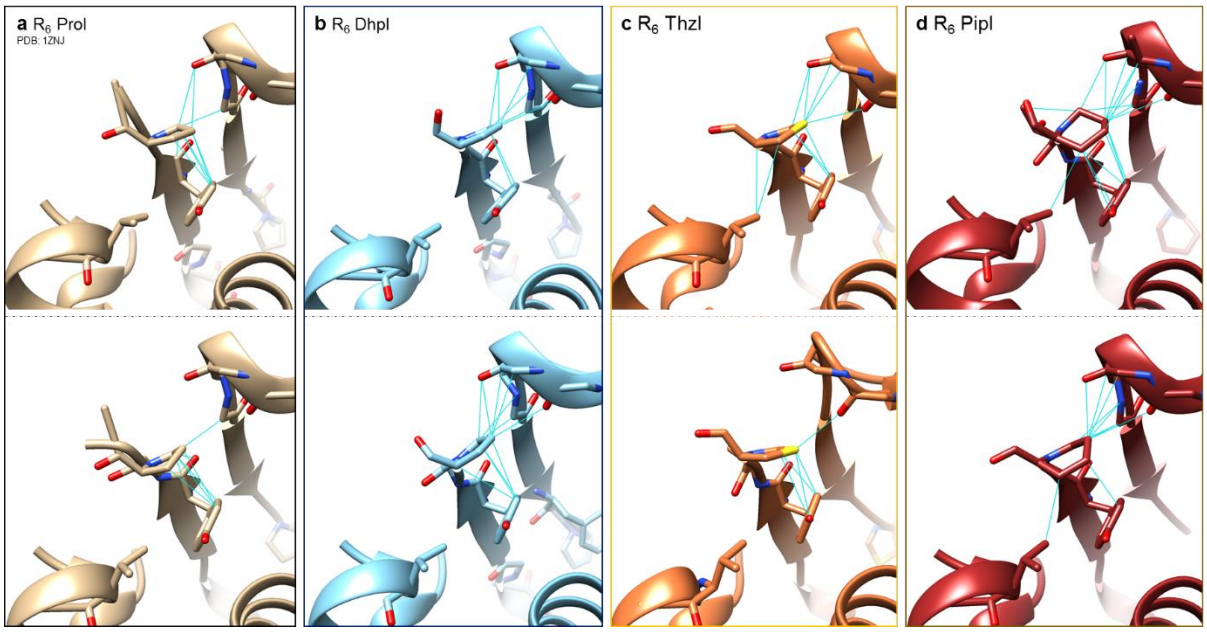
*T_2 crystal structures of ProI (**a** ,**b**) and AzeI (**c**). Geometric schematic of CH-π interaction shown in panel (**d**) for reference[31,32].*



We used the same crystallography software to also compute the network of vdW contacts within the R_6 structures (Figure 4.7). The results for the R_6 structures are nearly identical to the network from the T_2 structures (Figure 4.5d-f), except for the interaction between ValA3 and the α- and β-carbons at positon B28. This interaction is present in both monomers for PipI, one of two monomers for ThzI, and is absent in DhpI, suggesting that the appearance of this interaction may be a factor in tuning the rate of hexamer dissociation, since the rate from slow to fast follows the same order: PipI, ThzI, and DhpI.

Figure 4.7 | Van der Waals contacts at position B28 with neighboring atoms in the R-state.

*Crystal structures of R_6 insulins for ProI (**a**), DhpI (**b**), ThzI (**c**), and PipI (**d**). Images represent the two anti-parallel monomers. vdW contacts are represented by teal lines.*



Nature has evolved to use proline as one of twenty cAAs and the resulting protein sequence space has encompassed all the functions that are biologically needed; however, because protein therapeutics are subjected to synthetic conditions harsher than what has been biologically relevant, access to a larger chemical space may become necessary and

beneficial. Our results (Chapters 2-4) demonstrate the usefulness of ncAA mutagenesis, an approach that fuses the concepts of medicinal chemistry and protein design, and we anticipate that this protein engineering strategy will find increasing application in the design of antibody-drug conjugates, bispecific antibodies, and other novel protein therapeutics.

Materials and Methods

Materials. All canonical amino acids and (1,3)-thiazolidine-4-carboxylic acid (Thz) were purchased from Sigma. (3,4)-dehydro-L-proline (Dhp) and (*S*)-azetidine-2-carboxylic acid (Aze) was purchased from Bachem Americas. (*S*)-piperidine-2-carboxylic acid (Pip) was purchased from TCI America under the chemical name L-pipecolic acid. All solutions and buffers were made using double-distilled water (ddH_2O).

Strains and plasmids. The proinsulin (PI) gene with an *N*-terminal hexa-histidine tag (6xHIS), and flanked by *Eco*R1 and *Bam*H1 cut sites was ordered as a gBlock (Integrated DNA Technologies). Both the gBlock and vector pQE80L for IPTG-inducible expression were digested with *Eco*RI and *Bam*HI. Linearized vector pQE80L was dephosphorylated by alkaline phosphatase (NEB). Ligation of the digested PI gene and linearized vector yielded plasmid pQE80PI (to produce ProI). To make plasmid pQE80PI-proS (to produce FzpI, FypI and dFPI): Genomic DNA was extracted from *E. coli* strain DH10β using DNeasy Blood and Tissue Kit (Qiagen). Primers (Integrated DNA Technologies) were designed to amplify the *E. coli proS* gene, encoding prolyl-tRNA synthetase, under constitutive control of its endogenous promoter, from purified genomic DNA, and to append *Nhe*I and *Nco*I sites. The digested *proS* gene was then inserted into pQE80PI between transcription termination sites by ligation at *Nhe*I and *Nco*I restriction sites. Proline-auxotrophic *E. coli* strain CAG18515

and JT31 was obtained from the Coli Genetic Stock Center at Yale University. Prototrophic *E. coli* strain BL-21 was used for rich media expression of canonical insulins (ProI, AspI). Site-directed mutagenesis of pQE80PI at B28 was performed to make plasmid pQE80PI-asp, which differs from pQE80PI by three nucleotides that specify a single amino acid mutation to aspartic acid. All genes and plasmids were confirmed by DNA sequencing.

KS32 E. coli strain was made from parent *E. coli* strain JT31 using λ Red-mediated recombination[35]. JT31 was streaked from glycerol stock disk on an LB plate to obtain single colonies; a single colony was inoculated into an overnight culture of 2xYT media before a 200-fold dilution into the same medium the next morning (cultures were grown at 37°C, 200 RPM). At OD_{600} of approximately 0.4, cells were washed extensively in ice cold 10% glycerol to remove salts before resuspension in 10% glycerol at an OD_{600} of ~100. Plasmid pKD46 was transformed into electrocompetent JT31 cells by electroporation and plated on an LB Amp plate at 30°C to obtain single colonies. A single colony of JT31/pKD46 was made inoculated into an overnight culture of 2xYT media before a 200-fold dilution into the same medium the next morning (cultures here were grown at 30°C, 200 RPM). At an OD_{600} between 0.1 and 0.2, L-arabinose was added to the culture at a concentration of 10mM to induce expression of λ_{RED} recombinase. The cells were grown at 30°C until OD_{600} was between 0.4-0.6 before being washed and a final re-suspension with 10% glycerol at $0D_{600}$ of ~100. Ultramers flanking 50 bp upstream and downstream of the proA gene was ordered with IDT and used to perform PCR on plasmid pKD3 (flanking regions of proA with a FLP-CAT-FLP gene in-between); the PCR product was purified uising Zymo DNA concentrator spin columns transformed into electro-compentent JT31/pKD46 cells by electroporation

(Bio-Rad Electroporator). After rescuing of cells in SOC media for 1 h, cells were plated at 1:1 and 1:10 dilutions onto LB-CAM plates, and grown at 42°C to remove pKD46 plasmid. Single colonies were re-streaked onto LB-CAM and LB-AMP/CAM; colonies that grew on only LB-CAM plates were inoculated into 19aa and 20aa medium to check for auxotrophy. Genomic DNA from cells that grew in 20aa, but did not grow in 19aa medium, were purified using Qiagen's DNEasy Blood and Tissue Kit and sequenced to confirm removal of the proA gene. Colonies confirmed to have proA gene chromosomally removed were made chemically competent to uptake plasmid pCP20 for the removal of the FLP-CAT-FLP gene. The final strain *KS32* contains a scar in place of proA, which was also confirmed by sequencing.

Protein expression. Plasmids pQE80PI and pQE80PI-asp were transformed into BL21 cells and grown on ampicillin-selective agar plates. A single colony was used to inoculate 5 mL of Luria-Bertani (LB) medium and grown overnight; the resulting saturated culture was used to inoculate another 1 L of LB medium. All expression experiments were conducted at 37°C, 200 RPM in shake flasks (Fernbach 2.8 L flasks, Pyrex®). Each culture was induced with 1 mM IPTG at mid-exponential phase (OD_{600} ~0.8). For incorporation non-canonical prolines, pQE80PI-proS was transformed into CAG18515 (for Aze and Pip incorporation) or KS32 cells (for Dhp and Thz incorporation), which were grown on ampicillin-selective agar plates. To facilitate growth, a single colony was used to inoculate 25 mL of LB medium and the culture was grown overnight prior to dilution into 1 L of 1X M9, 20 aa medium (8.5 mM NaCl, 18.7 mM NH_4Cl, 22 mM KH_2PO_4, 47.8 mM Na_2HPO_4, 0.1 mM $CaCl_2$, 1 mM $MgSO_4$, 3 mg/L $FeSO_4$, 1 µg/L of trace metals (Cu^{2+}, Mn^{2+}, Zn^{2+}, MoO_4^{2-}), 35 mg/L thiamine hydrochloride, 10 mg/L

biotin, 20 mM D-glucose, 200 mg/L ampicillin with 50 mg/L of L-amino acids, each). At an appropriate cell density (OD_{600} ~0.8), the culture was subjected to a medium shift; briefly, cells were centrifuged and washed with saline prior to resuspension into 0.8 L of 1.25X M9, 19 aa (1X M9, 20 aa medium without L-proline). After cells were further incubated for 30 min to deplete intracellular proline, 200 mL of 5X additives (1.5 M NaCl, 2.5 mM Aze, Pip, Dhp, or Thz) was added to the culture. After another 15 min of incubation at 37°C to allow amino acid uptake prior to induction, IPTG was added to a final concentration of 1 mM. At the end of 2 h, cells were harvested by centrifugation and stored at -80°C until further use.

Cell lysis and refolding from inclusion bodies. Cells were thawed on the benchtop for 15 min prior to resuspension in lysis buffer (B-PER®, 0.5 mg/mL lysozyme, 50 U/mL benzonase nuclease). Cells were gently agitated at RT for 1 h prior to centrifugation (10 000 g, 10 min, RT); supernatant was discarded and the pellet was washed thrice: once with wash buffer (2 M urea, 20 mM Tris, 1% Triton X-100, pH 8.0) and twice with sterile ddH_2O; centrifugation followed each wash and the supernatant was discarded. The final washed pellet containing inclusion bodies (IBs, ~50% PI) was re-suspended in Ni-NTA binding buffer (8 M urea, 300 mM NaCl, 50 mM NaH_2PO_4, pH 8.0) overnight at 4°C or at RT for 2 h, both with gentle agitation. The suspension was centrifuged to remove insoluble debris; the remaining pellet was discarded and the supernatant was mixed with pre-equilibrated Ni-NTA resin (Qiagen) at RT for 1 h in order to purify PI from the IB fraction. Unbound proteins in the IB fraction were collected in the flow-through (FT), and the resin was washed with Ni-NTA wash buffer (8 M urea, 20 mM Tris base, 5 mM imidazole, pH 8.0) and Ni-NTA rinse buffer (8 M urea, 20 mM Tris base, pH 8.0) prior to stripping PI from the resin with Ni-NTA elution buffer (8 M

urea, 20 mM Tris base, pH 3.0). Fractions (IBs, FT, W, elution) were collected and run under reducing conditions on SDS-PAGE (Bis/Tris gels, Novex®); elution fractions containing PI were pooled and solution pH was adjusted to 9.6 with 6 N NaOH in preparation for oxidative sulfitolysis. Oxidative sulfitolysis was performed at RT for 4 h, with the addition of sodium sulfite and sodium tetrathionate (0.2 M Na_2SO_3, 0.02 M $Na_2S_4O_6$); the reaction was quenched by 10-fold dilution with ddH_2O. To isolate PI from the quenched solution, the pH was adjusted to between 3.5 and 4.5 by adding 6 N HCl dropwise; the solution became cloudy. The solution was centrifuged (10 000 g, 10 min, RT) and supernatant discarded. The PI pellet was then re-suspended in refolding buffer (0.3 M urea, 50 mM glycine, pH 10.6) and protein concentration was estimated by the bicinchoninic acid assay (BCA assay, Pierce®). The concentration of PI was adjusted to 0.5 mg/mL. Refolding was initiated by addition of β-mercaptoethanol to a final concentration of 0.5 mM and allowed to proceed at 12°C overnight with gentle agitation (New Brunswick® shaker, 100 RPM). Post-refolding, soluble PI was harvested by adjusting the pH of the solution to 4-5 by dropwise addition of 6 N HCl and by high speed centrifugation to remove insoluble proteins. The supernatant was adjusted to pH 8-8.5 by dropwise addition of 6 N NaOH and dialyzed against fresh PI dialysis buffer (7.5 mM sodium phosphate buffer, pH 8.0) at 4°C with five buffer changes to remove urea. The retentate (PI in dialysis buffer) was then lyophilized and subsequently stored at -80°C until further processing. Typical yields were 25-50 mg PI per L of culture (25-30 mg/L for non-canonical PI, 40-50 mg/L for canonical PI expression in rich media)

Proteolysis and chromatographic (HPLC) purification. The dry PI powder was re-dissolved in water to a final concentration of 5 mg/mL PI (final concentration of sodium phosphate

buffer is 100 mM, pH 8.0). Trypsin (Sigma-Aldrich) and carboxypeptidase-B (Worthington Biochemical) were added to final concentrations of 20 U/mL and 10 U/mL, respectively to initiate proteolytic cleavage. The PI/protease solution was incubated at 37°C for 2.5 h; proteolysis was quenched by addition of 0.1% trifluoroacetic acid (TFA) and dilute HCl to adjust the pH to 4. Matured insulin was purified by reversed phase high-performance liquid chromatography (HPLC) on a C_{18} column using a gradient mobile phase of 0.1% TFA in water (solvent, A) and 0.1% TFA in acetonitrile (ACN; solvent, B). Elution was carried from 0% B to 39% B with a gradient of 0.25% B per minute during peak elution. Fractions were collected and lyophilized, and the dry powder was re-suspended into 10 mM sodium phosphate, pH 8.0. Insulin-containing fractions were verified by matrix-assisted laser desorption/ionization-mass spectrometry (MALDI-MS; Voyager MALDI-TOF, Applied Biosystems) and SDS-PAGE to ensure identify and purity. Typical yields were 5-10 mg insulin per 100 mg PI. Fractions were stored at -80°C in 10 mM phosphate buffer, pH 8.0 until further use.

Verification of ncPro incorporation levels and maturation. A 30 µL aliquot of PI solution (8 M urea, 20 mM Tris, pH 8) was subjected to cysteine reduction and alkylation (5 mM DTT, 55°C, 20 min; 15 mM iodoacetamide, RT, 15 min, dark) prior to 10-fold dilution into 100 mM NH_4HCO_3, pH 8.0 (100 µL final volume). Peptide digestion was initiated with 0.6 µL of gluC stock solution (reconstituted at 0.5 µg/µL with ddH_2O, Promega) at 37°C for 2.5 h. The reaction was quenched by adding 10 µL of 5% TFA and immediately subjected to C_{18} ZipTip (Millipore) peptide purification and desalting according to the manufacturer's protocol. Peptides were eluted in 50% ACN, 0.1% TFA; the eluent was then diluted three-fold into

matrix solution (saturated α-cyanohydroxycinnamic acid in 50% ACN, 0.1% TFA) and analyzed by mass spectrometry (Voyager MALDI-TOF, Applied Biosystems). Incorporation levels were analyzed prior to and after refolding; incorporation percentage was calculated by comparing total AUC (area under the curve, arbitrary units) of the non-canonical peak with total AUC of its wild-type counterpart (1557 Da and 5808 Da, respectively). Maturation of AzeI, PipI, DhpI, or ThzI was analyzed after HPLC purification. TFA (1.6 µL, 5%) was added to 15 µL mature insulin solution (10 mM phosphate buffer pH 8.0) and subjected to C_{18} ZipTip (Millipore) peptide purification and desalting per the manufacturer's protocol. MALDI-MS conditions described above were used to confirm insulin maturation.

Circular Dichroism. Spectra were collected in a 1 cm quartz cuvette in 50 mM sodium phosphate buffer pH 8.0. Data were collected from 195 nm to 250 nm, with step size of 0.25 nm and averaging time of 1 s on a Model 410 Aviv Circular Dichroism Spectrophotometer; spectra were averaged over 3 repeat scans. A reference buffer spectrum was subtracted from the sample spectra for conversion to mean residue ellipticity. Insulin concentrations ranged from 3 µM to 250 µM.

Hexamer Dissociation Assay. Insulins were quantified by both UV absorbance (NanoDrop Lite, ThermoFisher) and BCA assay, and normalized to 125 µM insulin prior to dialysis against 50 mM Tris/perchlorate, 25 µM zinc sulfate, pH 8.0 overnight at 4°C using a D-tube dialyzer (Millipore Corp.) with MWCO of 3.5 kDa. Aliquots of dialyzed insulin solution were mixed with phenol to yield samples of the following composition: 100 µM insulin, 20 µM zinc sulfate, 100 mM phenol. Dissociation was initiated by addition of terpyridine (Sigma-Aldrich) to a final concentration of 0.3 mM from a 0.75 mM stock solution. A Varioskan

multimode plate reader (Thermo Scientific) was used to monitor absorbance at 334 nm. Kinetic runs were done at least in triplicate, and the data were fit to a mono-exponential function using Origin software. Post assay insulin samples were pooled and sample quality was determined by SDS-PAGE.

Fibrillation Assay. Insulin samples (60 μM in 10 mM phosphate, pH 8.0) were centrifuged at 22 000 g for 1 h immediately after addition of thioflavin T (ThT) (EMD Millipore) to a final concentration of 1 μM. Samples were continuously shaken at 960 RPM on a Varioskan multimode plate reader at 37°C, and fluorescence readings were recorded every 20 min for 48 h (excitation 444 nm, emission 485 nm). Assays were run in quadruplicate, in volumes of 200 μL in sealed (Perkin-Elmer), black, clear-bottom 96 well plates (Grenier BioOne).

Crystallographic Studies. Insulin crystals were obtained from sitting drop trays set using a Mosquito robot (TTP Labtech). Drops were set by mixing 0.4 μL insulin solution with 0.4 μL well solution. Cells were cryoprotected in a mother liquor containing 30% glycerol prior to looping and flash freezing in liquid nitrogen. Data were collected at SSRL beamline BL12-2 using a DECTRIS PILATUS 6M pixel detector. Initial indexing and scaling was performed with XDS; for some structures, data were re-scaled in alternative space groups using Aimless[36]. Initial phases were generated by molecular replacement in PHASER with either 3T2A (T_2 structures) or 1EV3 (R_6 structures)[37]. Structure refinement was carried out in Coot and Refmac5[38,39]. All distances and contacts were computed using UCSF Chimera Crystallography Software.

References

1. Brange, J. et al. Monomeric insulins obtained by protein engineering and their medical implications. *Nature* **333**(6174): 679-682 (1988).
2. Ciszak, E. et al. Role of C-terminal B-chain residues in insulin assembly: the structure of hexameric LysB28ProB29-human insulin. *Structure* **3**(6): 615-22 (1995).
3. Ho, B.K., Coutsias, E.A., Seok, C. & Dill, K.A. The flexibility in the proline ring couples to the protein backbone. *Protein Sci.* **14**(4): 1011-1018 (2005).
4. Reiersen, H. & Rees, A.R. The hunchback and its neighbours: proline as an environmental modulator. *Trends in Biochemical Sciences* **26**(11): 679-684 (2001).
5. Visiers, I., Braunheim, B.B. & Weinstein, H. Prokink: a protocol for numerical evaluation of helix distortions by proline. *Protein Eng.* **13**(9): 603-606 (2000).
6. S.P. Sansom, M. & Weinstein, H. Hinges, swivels and switches: the role of prolines in signalling via transmembrane α-helices. *Trends in Pharmacological Sciences* **21**(11): 445-451 (2000).
7. Zagari, A., Némethy, G. & Scheraga, H.A. The effect of the L-azetidine-2-carboxylic acid residue on protein conformation. I. Conformations of the residue and of dipeptides. *Biopolymers* **30**(9-10): 951-959 (1990).
8. Zagari, A., Némethy, G. & Heraga, H.A. The effect of the L-azetidine-2-carboxylic acid residue on protein conformation. II. Homopolymers and copolymers. *Biopolymers* **30**(9-10): 961-966 (1990).
9. Trotter, E.W., Berenfeld, L., Krause, S.A., Petsko, G.A. & Gray, J.V. Protein misfolding and temperature up-shift cause G(1) arrest via a common mechanism dependent on heat shock factor in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* **98**(13): 7313-7318 (2001).
10. Hayes, C.S., Bose, B. & Sauer, R.T. Proline residues at the C-terminus of nascent chains induce SsrA tagging during translation termination. *J. Biol. Chem.* **277**(37): 33825-33832 (2002).
11. Budisa, N. et al. 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. U. S. A.* **95**(2): 455-9 (1998).
12. Lin, L.N. & Brandts, J.F. Determination of cis-trans proline isomerization by trypsin proteolysis. Application to a model pentapeptide and to oxidized ribonuclease A. *Biochemistry* **22**(3): 553-559 (1983).
13. Kern, D., Schutkowski, M. & Drakenberg, T. Rotational barriers of cis/trans isomerization of proline analogues and their catalysis by cyclophilin. *JACS* **119**(36): 8403-8408 (1997).
14. Zagari, A., Némethy, G. & Scheraga, H.A. The effect of the L-azetidine-2-carboxylic acid residue on protein conformation. III. Collagen-like poly(tripeptide)s. *Biopolymers* **30**(9-10): 967-974 (1990).
15. Zagari, A., Palmer, K.A., Gibson, K.D., Némethy, G. & Scheraga, H.A. The effect of the L-azetidine-2-carboxylic acid residue on protein conformation. IV. Local substitutions in the collagen triple helix. *Biopolymers* **34**(1): 51-60 (1994).

16. Behre, J., Voigt, R., Althöfer, I. & Schuster, S. On the evolutionary significance of the size and planarity of the proline ring. *Naturwissenschaften* **99**(10): 789-799 (2012).
17. Berendes, R., Voges, D., Demange, P., Huber, R. & Burger, A. Structure-function analysis of the ion channel selectivity filter in human annexin V. *Science* **262**(5132): 427 (1993).
18. Choudhary, A., Pua, K.H. & Raines, R.T. Quantum mechanical origin of the conformational preferences of 4-thiaproline and its S-oxides. *Amino acids* **41**(1): 181-186 (2011).
19. Kubyshkin, V. & Budisa, N. cis–trans-Amide isomerism of the 3,4-dehydroproline residue, the 'unpuckered' proline. *Beilstein Journal of Organic Chemistry* **12**: 589-593 (2016).
20. Karzai, A.W., Roche, E.D. & Sauer, R.T. The SsrA-SmpB system for protein tagging, directed degradation and ribosome rescue. *Nat Struct Mol Biol* **7**(6): 449-455 (2000).
21. 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* **5**(7): 928-936 (2004).
22. Wood, J.M. Genetics of L-proline utilization in Escherichia coli. *J Bacteriol* (0021-9193 (Print))(1981).
23. Deutch, C.E. Oxidation of 3,4-dehydro-D-proline and other D-amino acid analogues by D-alanine dehydrogenase from Escherichia coli. *FEMS Microbiol Lett* **238**(2): 383-9 (2004).
24. Antolikova, E. et al. Non-equivalent role of inter- and intramolecular hydrogen bonds in the insulin dimer interface. *J. Biol. Chem.* **286**(42): 36968-77 (2011).
25. Brems, D.N. et al. Altering the association properties of insulin by amino acid replacement. *Protein Eng.* **5**(6): 527-533 (1992).
26. Pocker, Y. & Biswas, S.B. Self-association of insulin and the role of hydrophobic bonding: a thermodynamic model of insulin dimerization. *Biochemistry* **20**(15): 4354-4361 (1981).
27. Marshall, H., Venkat, M., Seng, N.S., Cahn, J. & Juers, D.H. The use of trimethylamine N-oxide as a primary precipitating agent and related methylamine osmolytes as cryoprotective agents for macromolecular crystallography. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **68**(Pt 1): 69-81 (2012).
28. Cowell, S.M., Lee, Y.S., Cain, J.P. & Hruby, V.J. Exploring Ramachandran and Chi Space: Conformationally Constrained Amino Acids and Peptides in the Design of Bioactive Polypeptide Ligands. *Current Medicinal Chemistry* **11**(21): 2785-98 (2004).
29. Li, A.J. & Nussinov, R. A set of van der Waals and coulombic radii of protein atoms for molecular and solvent‐accessible surface calculation, packing evaluation, and docking. *Proteins: Structure, Function, and Genetics* **32**(1): 111-127 (1998).
30. Kastritis, P.L. & Bonvin, A.M.J.J. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface* **10**(79)(2012).
31. Brandl, M., Weiss, M.S., Jabs, A., Sühnel, J. & Hilgenfeld, R. C-H···π-interactions in proteins. *J Mol Biol* **307**(1): 357-377 (2001).

32. Bhattacharyya, R. & Chakrabarti, P. Stereospecific Interactions of Proline Residues in Protein Structures and Complexes. *J Mol Biol* **331**(4): 925-940 (2003).
33. Biedermannova, L., E. Riley, K., Berka, K., Hobza, P. & Vondrasek, J. Another role of proline: stabilization interactions in proteins and protein complexes concerning proline and tryptophane. *Physical Chemistry Chemical Physics* **10**(42): 6350-6359 (2008).
34. Zondlo, N.J. Aromatic-Proline Interactions: Electronically Tunable CH/π Interactions. *Accounts of Chemical Research* **46**(4): 1039-1049 (2013).
35. Datsenko, K.A. & Wanner, B.L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci. U. S. A.* **97**(12): 6640-5 (2000).
36. Winn, M.D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **67**(Pt 4): 235-242 (2011).
37. McCoy, A.J. et al. Phaser crystallographic software. *J Appl Crystallogr.* **40**(Pt 4): 658-674 (2007).
38. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **66**(Pt 4): 486-501 (2010).
39. Murshudov, G.N., Vagin, A.A. & Dodson, E.J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **53**(3): 240-255 (1997).

Acknowledgements

We thank J. T. Kaiser and P. Nikolovski of the Molecular Observatory at Caltech, and S. Russi and the scientific staff of Beamline 12-2 at the Stanford Synchrotron Radiation Laboratory for assistance. We thank S. Virgil of the Chemical Catalysis Center and M. Shahgholi of the Mass Spectrometry Facility at Caltech for their assistance. We thank W. Glenn for editing this chapter.

This work was done in collaboration with Seth Lieblich. S.L. performed experiments for insulin maturation and HPLC purification, sample preparation for obtaining circular dichroism spectra and solving crystal structures of insulin.

Chapter 5 – Future avenues for insulin engineering: Further development of an expanded genetic code

Abstract

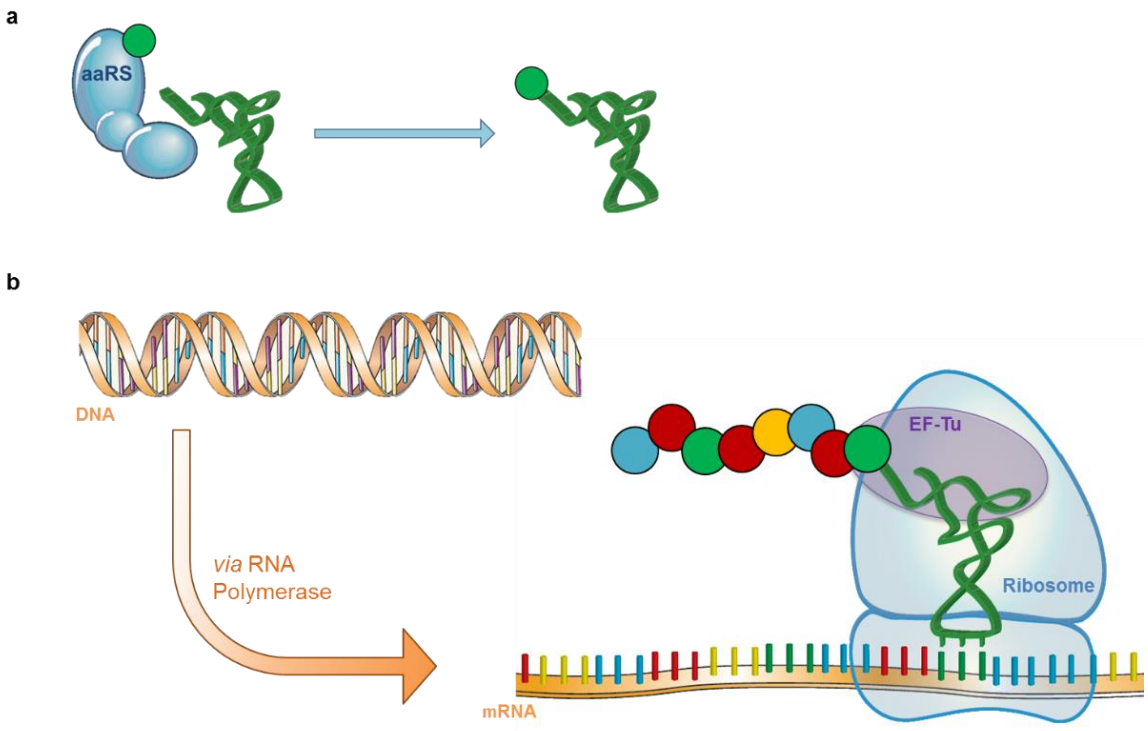
Non-canonical amino acids (ncAAs) have provided great insights to the field of protein engineering. However, in many circumstances, the quantity of protein obtained is an important factor, and current site-specific and non-bacterial expression systems typically result in low protein yields. Here we describe a general and facile screening method for the discovery of mutant aminoacyl-tRNA synthetases for the incorporation of new ncAAs using *E. coli* as an expression host.

Introduction

Engineering aminoacyl-tRNA synthetases. Nature is limited in protein sequence space by the twenty cAAs and thus, natural evolution[1] has evolved proteins to function and have properties as far as it is biologically needed. Protein therapeutics, such as insulin, are subjected to conditions harsher than the endogenous conditions in which a protein was evolved to function in. It can be very possible that certain proteins will never evolve to perform a certain function or function optimally under synthetic conditions because there has never been any selective pressure to do so. Therefore, in this respect, chemically distinct amino acids are beneficial and can lead to new and improved protein functions and properties that cannot be (easily) accessed in the canonical protein sequence space[2,3].

Figure 5.1 | Scheme depicting protein translation.

***a**, Aminoacylation reaction catalyzed by aminoacyl-tRNA synthetase (aaRS). **b**, Central dogma of molecular biology. Figure adapted from scheme previously described[4].*



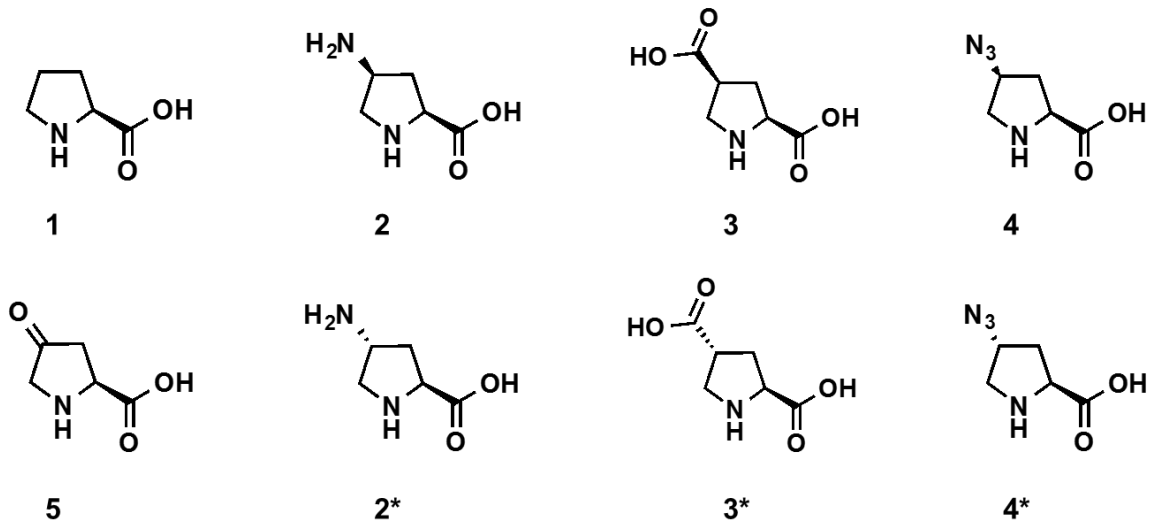
The process for protein translation begins with the transcription of DNA into messenger RNA (mRNA). The ribosome is then responsible for translating the mRNA into polypeptide, which will fold and may undergo post-translational processes to form the final protein. Translation involves two key components: the ribosome complex and aminoacylated tRNAs (Figure 5.1a). Because the ribosome complex recognizes the RNA portion of the aminoacylated tRNA, there is an opportunity to take advantage of the aminoacylation reaction to introduce non-canonical amino acids (ncAAs) into the endogenous translation process: if an ncAA can be ligated to its cognate tRNA by an aminoacyl-tRNA synthetase (aaRS), then the ribosomal complex is typically promiscuous enough to allow for translation to proceed. Therefore, it is not a surprise that all current biological approaches for the incorporation of ncAA involve aaRS and their cognate tRNAs. It should be noted that the ribosomal complex does have elongation factors (e.g. EF-Tu) that must be engineered to allow incorporation of certain ncAAs; however, there are only a few such reported cases and all involve large, bulky amino acids (e.g. phosphoserine, selenocysteine[5]).

Many aaRS have been discovered through high-throughput screening methods for the incorporation of non-canonical amino acids (ncAAs). These engineered aaRS have allowed for both site- and residue-specific incorporation of ncAA for a variety of protein-related applications (e.g., cell imaging, conjugation, and engineering protein therapeutics). A general screening method to develop aaRS for site-specific incorporation[6-8] of ncAAs has been reported. However, there is no general screening method for residue-specific incorporation of ncAAs; instead, each of the published methods is specific to an aaRS for a given amino acid. If it is desired that a ncAA be incorporated residue-specifically and the

endogenous aaRS cannot accommodate the new amino acid, then the aaRS must be adapted. For example, methods for engineering mutant methionyl-tRNA synthetases (MetRS) have been developed because of interest in using methionine analogs (ncMet) in proteomics experiments (BONCAT). For this purpose, GFP_{1Met} (and related variants) was evolved so its fluorescence capability would be insensitive to methionine replacement to directly correlate fluorescence with ncMet incorporation. Through fluorescence-activated cell sorting[9-11], it was then possible to isolate the high fluorescing cells expressing mutant MetRS capable of incorporating the desired methionine analog of interest.

Figure 5.2 | Additional proline analogs for study of insulin biophysics.

Compound **1***: L-proline; Compound* **2***: (2S,4S)-4-amino-L-proline (Nzp); Compound* **3***: (2S,4S)-4-carboxy-L-proline (Czp); Compound* **4***: (2S,4S)-4-azido-L-proline (Azp); Compound* **5***: 4-oxo-L-proline (Kep); * denotes (2S,4R) variant*



Proline. Investigating proline analogs is important because studies in the literature show that proline has a critical role in protein folding[12], and can influence oligomeric and aggregation behavior[13,14], act as the gatekeeper allow molecules to flow through ion channels or receptors[15], and influence other significant biological functions[16].

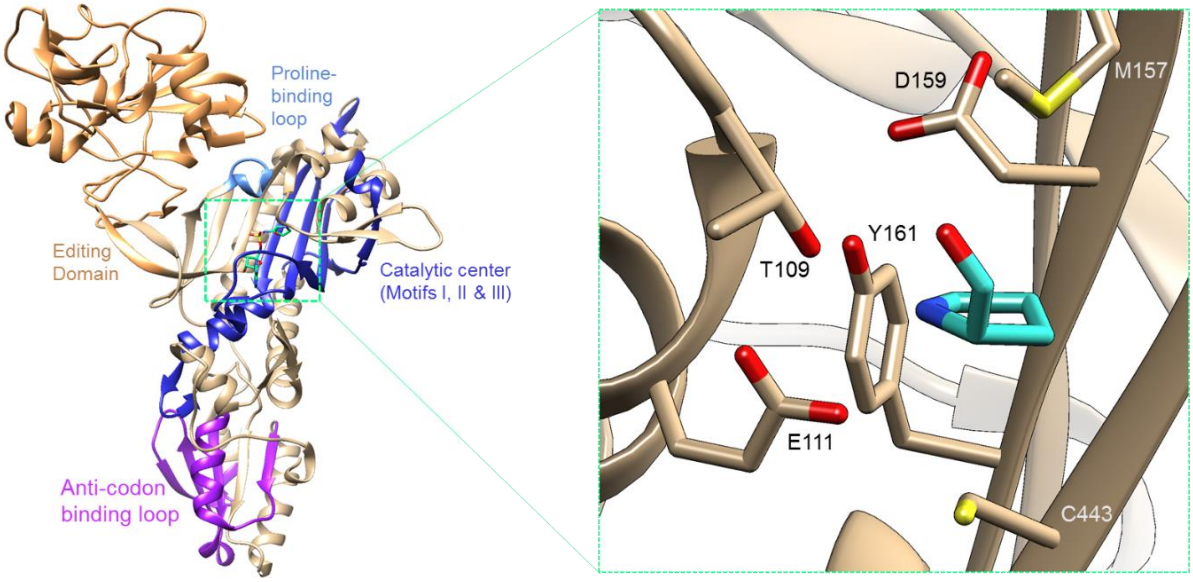
Chapters 2-4 discussed the use of proline analogs to modulate the biophysical properties of insulin. One of these analogs, (*2S,4S*)-4-hydroxy-L-proline (Hzp), when in place of the proline residue at position B28, yielded an improved insulin (HzpI), with fast dissociation kinetics and enhanced resistance to fibril formation. Crystallographic analysis revealed that this may be due to the appearance of a novel hydrogen bond. To build on this hypothesis, we sought to utilize a set of analogs with varying polarity and propensity as a hydrogen bond donor (Figure 5.2), and conformational preferences[17-20]. There is no literature precedent for non-synthetic methods of incorporation for many ncPros; to this end, it would be useful to develop a recombinant expression system that can do so.

Prolyl-tRNA synthetase. There are two major classes of aaRS: class I synthetases are characterized by a Rossman domain in the active site and approach tRNA from the minor groove side, while class II synthetases have an antiparallel β-fold in the active site and approach tRNA from the major groove side[21]. Prolyl-tRNA synthetase (ProRS) is considered a class II synthetase[22]. Regardless of categorization, each aaRS serves the larger function of charging an amino acid onto the 3'-hydroxyl group of its corresponding tRNA, which requires holding the amino acid in proximity to its cognate tRNA and allowing the aminoacylation reaction to proceed[21]. Since each aaRS has evolved to recognize the size, shape, and characteristics (e.g., hydrophobicity, charge, polarity) of its cAA[23,24], the more a ncAA deviates from its analogous cAA in terms of these properties, the less likely it is the endogenous aaRS will be able catalyze the aminoacylation reaction. Therefore, by engineering the amino acid binding pocket without significantly perturbing any other interactions, the aaRS can catalyze the reaction with the same tRNA, but with a ncAA in

place of a cAA[25]. Unfortunately, there is no published crystal structure of *E. coli* ProRS (*Ec* ProRS), but considering that the twenty cAAs, and tRNAs are conserved across species[26], we referenced the crystal structure of the *Enterococcus faecalis* ProRS (*Ef* ProRS, 46% sequence identity to *Ec* ProRS). The residues that are within close proximity to proline in the binding pocket of ProRS represent a good starting point for deciding which residues to mutate to switch the synthetase's selectivity and activity towards other proline analogs (Figure 5.3). Mutations at several of these residues, particularly Cys443[27], can result in a more promiscuous synthetase.

Figure 5.3 | Prolyl-tRNA synthetase amino acid binding pocket

On left, Structure of the prokaryotic prolyl-tRNA synthetase (from Enterococcus faecalis ProRS*, PDB: 2J3L) with critical regions labeled and color coded[28,29]. Inset: Close-up representation of proline (shown in teal) in the proline binding pocket. Residue numbers correspond to the analogous residue for E. coli* ProRS*.*



Many studies[30-34] on protein evolution suggest that it is difficult to predict beneficial mutations that can enhance enzyme activity or protein stability. In some cases,[35,36] the

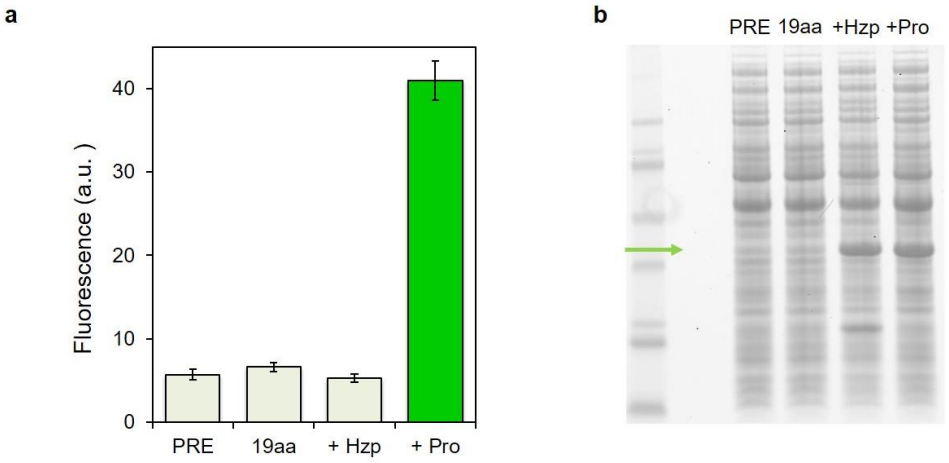
beneficial mutations found were not anywhere near the known active site; therefore, it is likely that synthetase engineering, beyond rational design of mutations of the binding residues, will require random mutagenesis. A high-throughput screening platform will be needed to search through millions of mutants to find a small subset containing beneficial mutations for increased incorporation of a particular ncPro.

Results and Discussion

Global replacement of proline residues can render full-length GFP non-fluorescent. Previous work with GFP_{1Met} and its variants demonstrated success in discovering new mutant MetRS for residue-specific incorporation of methionine analogs[10,11]. There is evidence in the literature for residue-specific incorporation of proline analogs (i.e. fluoroprolines[37], (3,4)-dehydro-L-proline[38]) in GFP without abolishing the protein's fluorescence capability.

Figure 5.4 | Residue-specific replacement of proline residues in GFP_{1Met}

***a**, Normalized fluorescence of cultures (n = 3) before (PRE) and after induction in minimal medium supplemented with 19 amino acids with the addition of either nothing (19aa), (2S,4S)-4-hydroxy-L-proline (+Hzp) or L-proline (+Pro). **b**, SDS-PAGE of cell lysates. Green arrow denotes expected location of GFP_{1Met} protein band.*



We first verified whether GFP fluorescence can be retained with Hzp globally replacing all Pro residues; Hzp contains a hydrophilic substituent group, which will provide a larger perturbation than the proline analogs previously incorporated into GFP from the literature. Preliminary studies on the incorporation of Hzp into GFP_{1Met} yielded a non-fluorescent protein (Figure 5.4), indicating fluorescence cannot be correlated with the incorporation of proline analogs and GFP_{1Met} cannot be used as to screen for new ProRS for the incorporation of proline analogs. One strategy would involve evolving GFP to become insensitive to proline replacement (e.g. remove all prolines in GFP or place proline residues at permissive locations); this has been done successfully with methionine and tryptophan[39]. However, bioinformatic analysis of fluorescent proteins through comparison of sequence homology suggest there are five conserved prolines (J. Cahn and S. Lieblich, personal communication, October 31, 2016), indicating that evolving GFP to remove proline residues may prove difficult or perhaps even impossible. An alternative approach may be to utilize a dual protein-complementation system to minimize the destabilizing effects resulting from global replacement of a cAA. In the past decade, many labs have demonstrated the utility of split-GFP reporter systems for investigating protein-protein interactions[40,41], screening protein solubility[42], mapping endogenous protein topology[43], and imaging cellular behavior[44,45]. We sought to adapt the GFP(1-10)/11 split GFP reporter system into a high-throughput screening platform for the discovery of mutant prolyl-tRNA synthetases that will enable residue-specific incorporation of ncPros into recombinant proteins in *E. coli*.

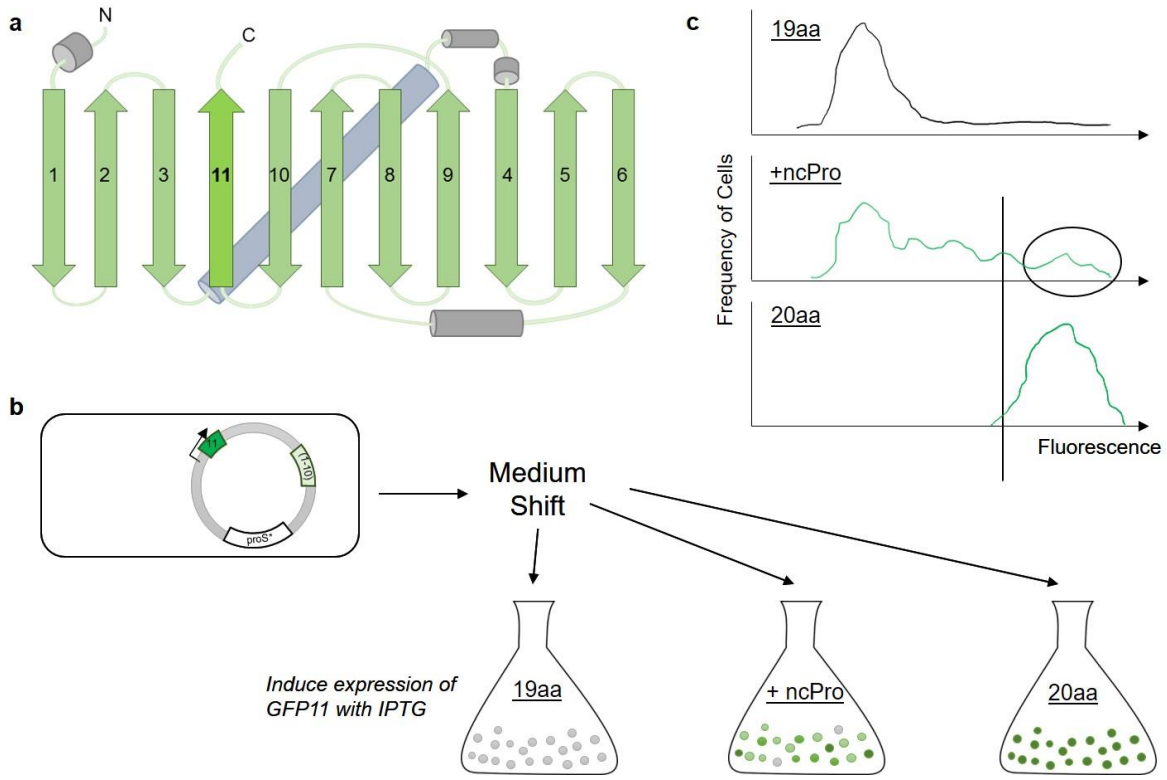
Split-GFP complementation in live bacterial cells. We chose to proceed with the split-GFP system to circumvent rendering GFP non-fluorescent. GFP is comprised of two parts, β-

sheet strands 1-10 and 11 (GFP(1-10) and GFP11), that are non-fluorescent independently and need to associate for split-GFP to become fluorescent (Figure 5.5a). Some studies in the literature[43,46] have appended large, globular proteins to either the N- or C-terminus of GFP11 and found that the two parts, GFP(1-10) and GFP11, of split-GFP can complement and fluoresce accordingly; therefore, the insertion of prolines to the N-terminus of GFP11 is unlikely to interfere with fluorescence output. Our goal is to grow cultures (containing a library of ProRS) with GFP(1-10) constitutively expressed before performing a medium shift at mid-growth phase to 19aa; after the depletion step (to remove excess intracellular proline), we will add the ncPro and induce expression of GFP11 (Figure 5.5b). Cells containing mutant synthetases with improved activity towards ncPro will be able to express GFP11, which will then associate with the existing pool of GFP(1-10) protein and fluoresce; these mutants will then be sorted and sequenced for further validation experiments (Figure 5.5c). Our initial design was a one-plasmid system, adapted from plasmid pQE80PI-proS (from Chapters 2-4), with polycistronic expression of the prolyl-tRNA synthetase and GFP(1-10), each gene with its own ribosomal binding site (RBS) for independent initiation of translation (Figure 5.6a-*i*). We inserted a proline at the N-terminus of GFP11 to correlate proline or proline analog incorporation with GFP11 expression and subsequent fluorescence output. Initial expression and fluorescence experiments found that the dynamic range between no induction or induction in 19aa (only GFP(1-10) present), and induction in 20aa (19aa + Pro) was two-fold (Figure 5.6a-*ii*); this may be due to the low levels of GFP(1-10) that are undetectable with colloidal blue staining (Figure 5.6a-*iii*). We could not see baseline

separation between the distributions of fluorescent and non-fluorescence cell populations using a flow cytometer (data not shown).

Figure 5.5 | Scheme for split-GFP complementation and incorporation.

***a**, Homologous fluorescent protein topology map adapted from scheme previously described[47]. Label for β sheet strand 11 (*β11*) is bolded. **b,** Scheme for split-GFP complementation and incorporation of ncPros. Cells harboring plasmid(s) containing constitutively expressed GFP(1-10) and ProRS are grown to mid-exponential phase before being subjected to a medium shift. GFP11 expression is induced thereafter in 19aa media with the addition of nothing (19aa), ncPro or Pro (20aa). **c,** Pictorial representation, not actual data, of fluorescence distributions for cell populations subjected to 19aa, 19aa + ncPro (+ncPro), or 19aa + Pro (20aa) expression conditions.*

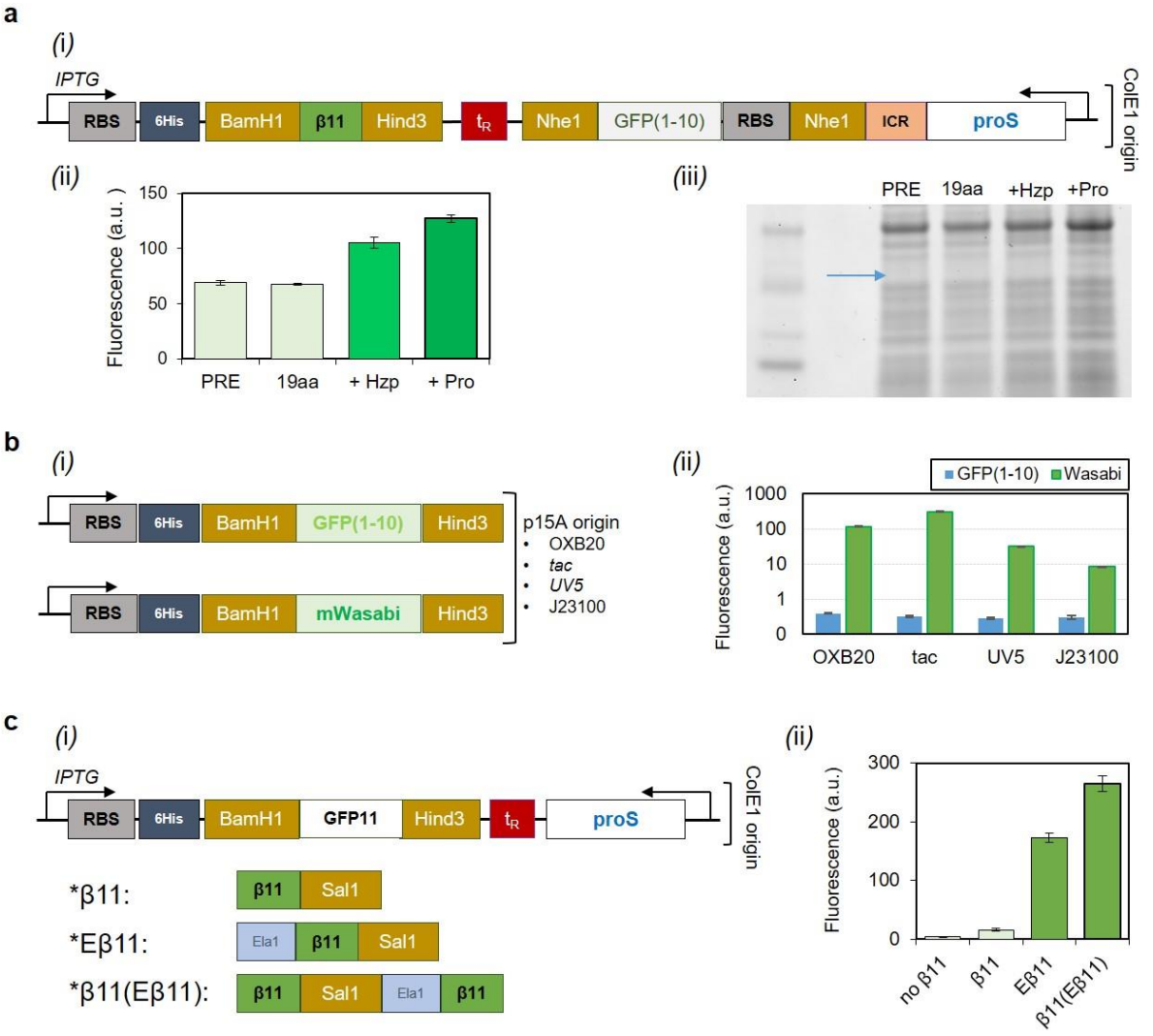


Selection of constitutive promoters for GFP(1-10) expression. We chose to place GFP(1-10) on a separate plasmid under its own promoter[48,49] (plasmid from Figure 5.6a-*i* is near 7 kb in length). Selection of promoter for GFP(1-10) expression was done using plasmids shown in Figure 5.6b-i with mWasabi (a GFP variant) a reporter protein; p15a is an origin of replication that is associated with medium copy numbers. We found that the *tac* promoter

had the best signal-to-noise ratio (Figure 5.6b-ii) when comparing mWasabi fluorescence to GFP(1-10) fluorescence . Plasmid p*tac*_GFP(1-10) was combined with GFP11 constructs (Figure 5.6c) and transformed into *E. coli* for *in vivo* expression and complementation.

Figure 5.6 | Split-GFP plasmids for screening mutant prolyl-tRNA synthetases

a, *Single plasmid (ColE1 origin) containing: 1) β11 under inducible lac promoter, 2) GFP(1-10) constitutively expressed with 3) gene encoding for ProRS under its endogenous promoter. (ii) Fluorescence levels from GFP(1-10)/β11 expression (n=3). (iii) SDS-PAGE of cell lysates; blue arrow indicating expected band size of GFP(1-10); absence of band indicates extremely low-level expression.* ***b,*** (i) *Plasmids used to select ideal promoter for enhancing GFP(1-10) expression. Promoters derived from* lac *had lacI binding site omitted to maintain constitutive expression. (ii) Fluorescence levels comparing signal (Wasabi, positive control) to noise (GFP(1-10), negative control) ratios.* ***c,*** (i) *GFP11 variants; E is the elastin sequence* $(VPGAG)_2VPGEG(VPGAG)_2$. (ii) *Fluorescence levels from cells containing ptac_GFP(1-10) and a GFP11 plasmid in 20aa medium, post-IPTG induction (n=3).*



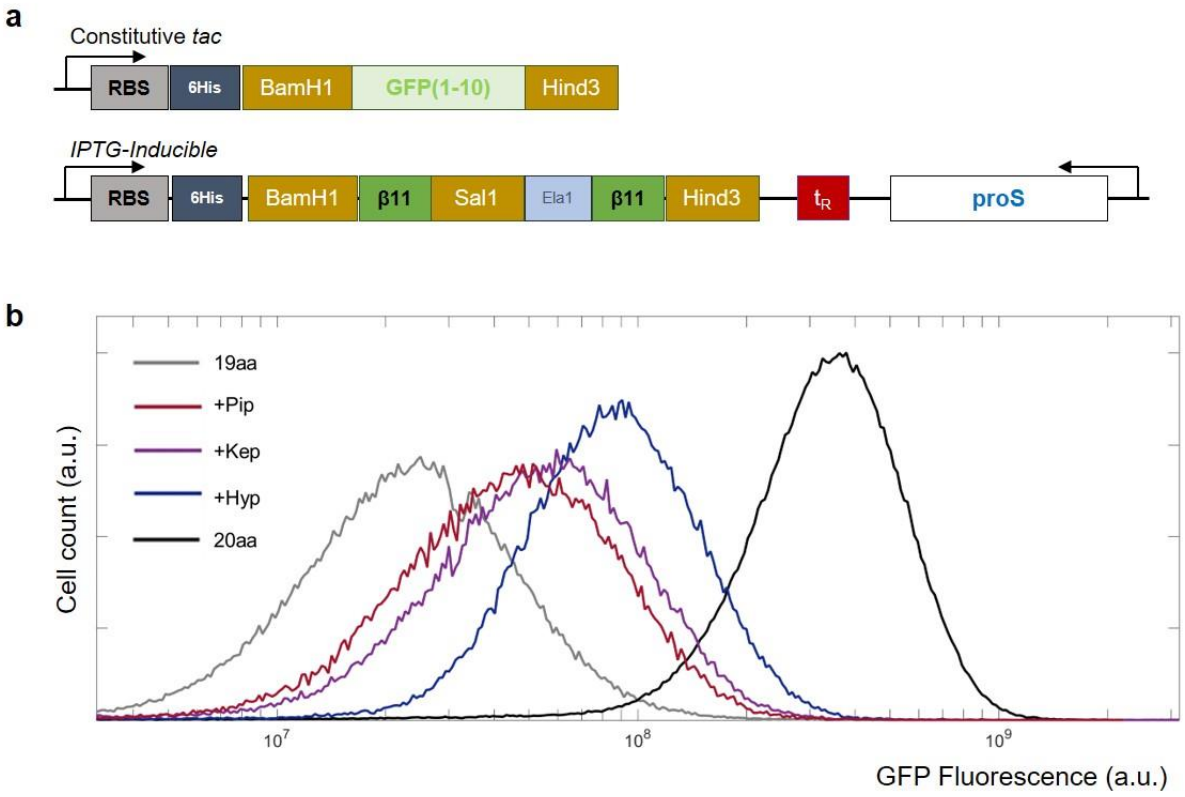
Design of GFP11 multi-blocks for enhancing split-GFP fluorescence. The overall GFP11 constructs followed the design of pQE80PI-proS plasmid constructs from Chapters 2-4; this was done with consideration for further insulin (or other therapeutic protein) engineering since we had previously obtained good expression yields and high incorporation percentages. We first needed to determine that plasmid design and expression conditions could facilitate re-assembly of the split GFP fragments and subsequent fluorescence readout. We designed three GFP11 constructs: β11, β11 with an N-terminal elastin[50-52] (Eβ11) to aid in expression, linker flexibility, and selectivity for proline incorporation (Eβ11 polypeptide sequence contains 7 Pro residues compared to 1 Pro residue in β11), and two β11s in tandem, with an elastin linker in-between (β11(Eβ11), containing 8 Pro residues in total). The latter was designed to increase fluorescence output based on evidence in the literature demonstrating that the more β11s in tandem correlates with higher fluorescence output[44]. Co-expression of GFP11 variants and GFP(1-10) with 20 cAAs determined that β11(Eβ11) yielded the highest average fluorescence output (Figure 5.6c-ii).

Plasmids, shown in Figure 5.7a, were transformed in proline-auxotrophic *E. coli* strain CAG18515 and grown in minimal media, supplemented with 20aa until mid-exponential phase. The cells were then shifted to minimal media, supplemented with 19aa for depletion before adding either nothing (19aa), L-pipecolic acid (Pip), 4-oxo-L-proline (Kep), (2S,4R)-hydroxy-L-proline (Hyp) or L-proline (20aa), followed by the addition of IPTG to induce GFP11 expression. Since the cells were constitutively expressing the wild-type synthetase, and based on previous experiments, we expected incorporation percentages to be: undetectable for 19aa and Pip, 30% for Keto, 90% for Hyp, and 100% for Pro. We

observe this trend in our fluorescence data (Figure 5.7b). In addition, we see good peak separation between our negative (19aa) and positive (+Pro) controls and thus, we elected to move forward with our screening system.

Figure 5.7 | Split-GFP system for screening prolyl-tRNA synthetase library.

__a__, Plasmids used for screening purposes. __b__, Fluorescence distribution of cell populations with plasmids in (__a__). Cultures were grown in 20aa (for GFP(1-10) constituitive expression) before medium shifted to 19aa, +Pip, +Keto, +Hyp, and +Pro (20aa) for IPTG-induction of GFP11 expression. Flow cytometry distributions are normalized to an area of 1 a.u.



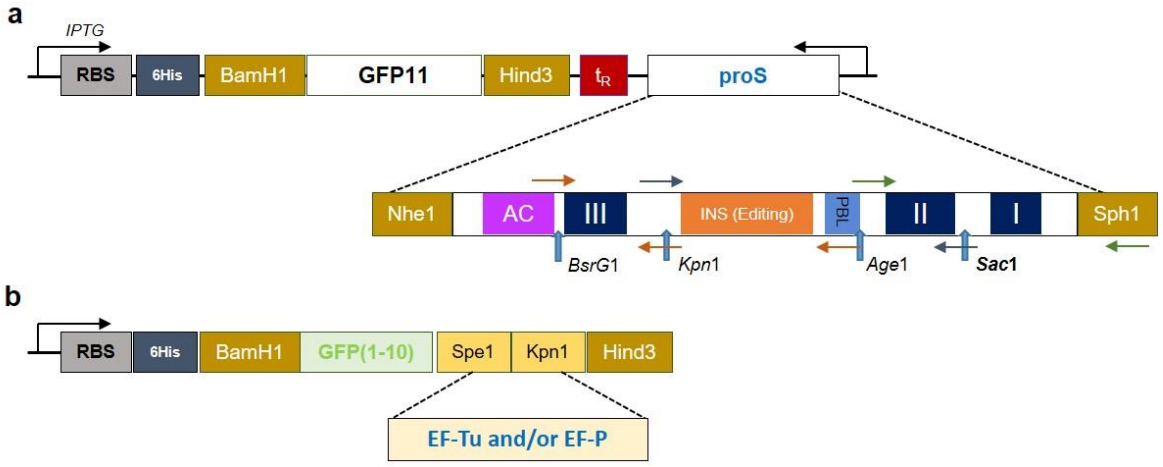
Ongoing Work and Future Implications

<u>Construction and preliminary screening of prolyl-tRNA synthetase libraries</u>. Plasmid pQE80(GFP11 inserts)-*proS was designed to be modular. Motifs I, II and III are flanked by unique restriction enzyme recognition sites to allow for site-saturation mutagenesis of active site residues. In addition, each domain of the synthetase is flanked by unique

restriction enzyme recognition sites to allow for focused random mutagenesis of specific segments (i.e. the editing domain).

Figure 5.8 | Scheme for library construction for incorporating proline analogs.

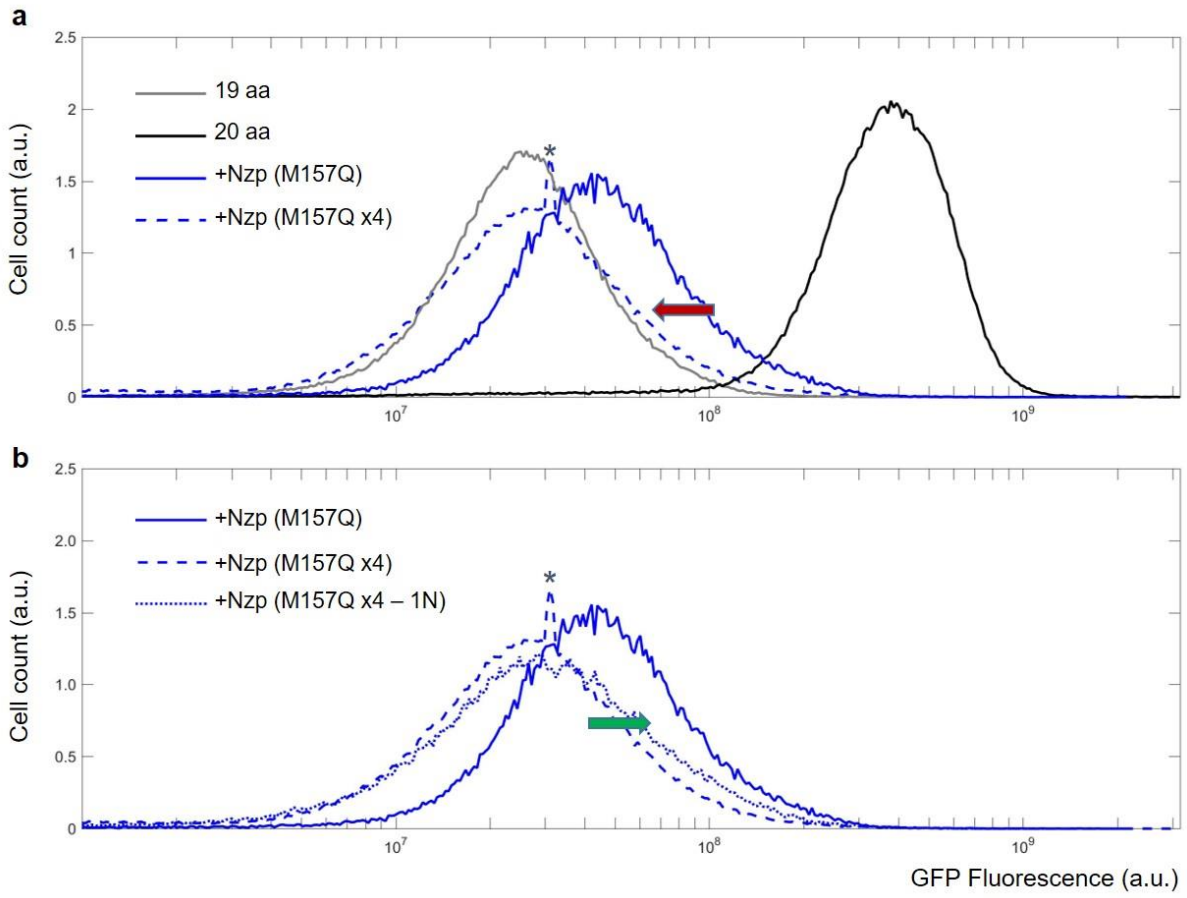
***a**, Plasmid design for constructing synthetase libraries. ProRS sections are presented in the close-up depication of the proS gene fragment: I, II & III = aminoacylation domain (navy blue); PBL = proline binding domain (pink); INS = ProRS editing domain (orange); AC = anti-codon binding loop domain (purple). Arrows denote fragments of proS for error-prone library construction. **b**, Plasmid design for constructing either EF-Tu or EF-P libraries.*



Site-saturation mutagenesis of residues in the proline binding pocket. Previous work on incorporating Nzp using plasmid pQE80PI-proS with mutations at methionine 157 (M157, M157D, M157E, M157N and M157Q) found incorporation levels varied between 0 to approximately 60%. Using the parent synthetase ProRS (M157Q), which results in an incorporation percentage of 60% for Nzp, we constructed the starting library (M157Q x4) by randomizing residues 158-161 with NNK codons (N = A, T, C or G; K = C or T) since residues D159 and Y161 are in the binding pocket (Figure 5.3). We choose NNK codons because 2 of 3 stop codons are eliminated from the possible codons. We obtained 10^7 transformants during both the cloning and expression transformation stages.

Figure 5.9 | Cell distributions for naïve library (site-saturation mutagenesis).

Cultures for flow cytometry were grown to mid-exponential phase for medium shift, followed by a brief depletion step and the addition of either nothing (19aa), (2S,4S)-4-amino-L-proline (+Nzp) or proline (20aa), prior to IPTG-induction of GFP11 expression. Cell distributions are normalized to an area of 1 a.u. ***a****, Fluorescence distribution of cell populations for the parent synthetase (M157Q) and naïve library (M157Q x4); 19aa and 20aa samples show as negative (non-fluorescent) and positive (fluorescent) controls.* ***b****, Fluorescence distribution of cell populations expressing GFP11 in the presence of Nzp for the parent synthetase (M157Q), naïve library (M157Q x4), top 0.5% sorted cells of the naïve library (M157Q x4-1N). *represents an anomalous peak from the flow cytometer*



The distribution for cells expressing GFP11 (using CAG18515 cells containing the plasmids shown in Figure 5.7a) in the presence of Nzp for the naïve library (M157Q x4) is less fluorescent than the distribution for cells expressing single mutant ProRS (M157Q). This indicates that site-saturation mutagenesis at residues 158-161 (D159 and Y161 are represented in Figure 5.3) primarily leads to mutant synthetases that are less active towards

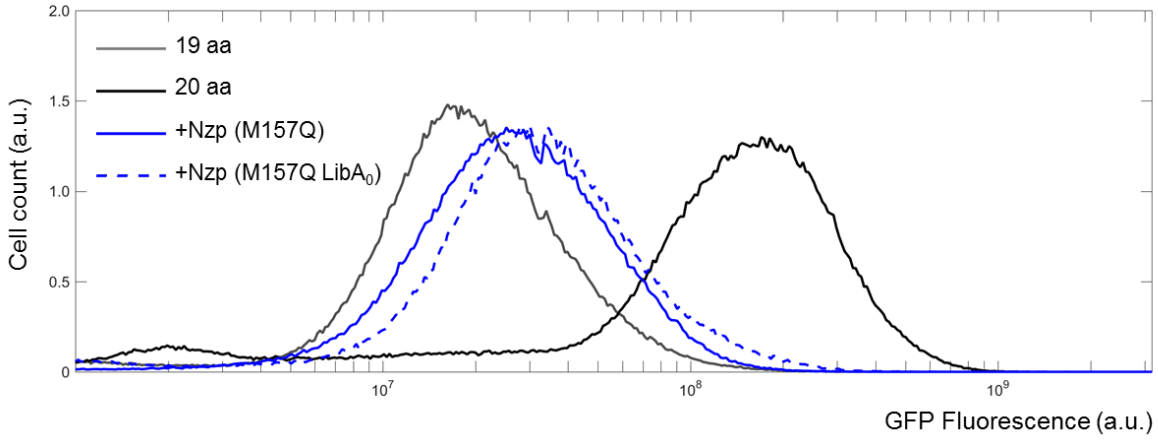
Nzp than the parent mutant synthetase (Figure 5.9a; red arrow denotes direction of mean fluorescence from parent to naïve library). One possible scenario is that further mutations of residues in the binding pocket of ProRS (M157Q) is unlikely to result in enhanced activity or selectivity towards Nzp. Nonetheless, we sorted the top 0.5% fluorescing cells from the naïve library to isolate library M157Q x4-1N. The mean fluorescence of the distribution for cells expressing GFP11 in the presence of Nzp for the sorted library (M157Q x4-1N) is slightly shifted to the right (Figure 5.9b; green arrow denotes direction from lower to higher fluorescence) compared to the naïve library (M157Q x4); however, this shift is minor and suggests the naïve library may not be the ideal starting point. In addition, the loss in mean fluorescence, comparable broadness of the fluorescence distribution, and similar number and fluorescence of top fluorescing cells between the naïve library and the parent synthetase are also indicators that it will be difficult to find a more active synthetase than the parent in this naïve library (M157Q x4).

Error-prone library. The design of the proS gene fragment allows us to further diversify prolyl-tRNA synthetase libraries, while avoiding mutations in important consensus domains (i.e. anti-codon binding loop; Figure 5.3, colored purple), by focusing mutations in regions (i.e. motifs I, II & III, editing domain and PBL) that are more likely to be important for switching synthetase selectivity and enhancing aminoacylation activity for proline analogs. Based on the results from library M157Q x4, we performed random mutagenesis on the gene segment encoding for PBL, editing domain and motif III of the parent synthetase ProRS (M157Q) to create the naïve error-prone library (M157Q libA_0). Using the same

transformation and expression protocols, we subjected M157Q libA$_0$ to GFP11 expression in the presence of Nzp (Figure 5.10).

Figure 5.10 | Cell distributions for naive library (random mutagenesis).

Cultures for flow cytometry were grown to mid-exponential phase for medium shift, followed by a brief depletion step and the addition of either nothing (19aa), (2S,4S)-4-amino-L-proline (+Nzp) or proline (20aa), prior to IPTG-induction of GFP11 expression. Cell distributions are normalized to an area of 1 a.u. Fluorescence distribution of cell populations for the parent synthetase (M157Q) and naïve library (M157Q LibA$_0$); 19aa and 20aa samples show as negative (non-fluorescent) and positive (fluorescent) controls.



The mean fluorescence of the distribution for cells expressing GFP11 in the presence of Nzp for the naïve error-prone library (M157Q LibA$_0$) does not display the loss in mean fluorescence compared to the parent synthetase (Figure 5.10) that we saw for the M157Q x4 library (Figure 5.9). In particular, at the right tail of the distributions of the cells expressing GFP11 in the presence of Nzp, the number of top fluorescing cells for the M157Q LibA$_0$ library exceeds that of the parent synthetase; although this result is preliminary, it may indicate that it may be possible to find a mutant synthetase with enhanced activity and/or selectivity towards Nzp compared to the parent ProRS (M157Q) mutant in the M157Q LibA$_0$ library.

Library construction for other translational machinery. There exists evidence in the literature that the overexpression of elongation factors (EF-Tu[53] and EF-P[54]) is necessary for incorporating ncAAs. For some ncAAs, a mutant EF-Tu is required[5,55,56] for translation to proceed. The role of EF-Tu is essential during the elongation[57] step of translation. Our split-GFP screen can be adapted to find mutant elongation factors by simply inserting an EF-Tu or EF-P library downstream of GFP(1-10) in the p*tac* plasmid (Figure 5.8b). In addition, expression levels of EF-Tu (or EF-P) can be optimized by using the same protocols used for increasing GFP(1-10) expression (Figure 5.6).

Continued screening of prolyl-tRNA synthetase libraries. Our interest in incorporating the proline analogs in Figure 5.2 is to further engineer insulin. In Chapter 2, we describe a novel hydrogen bond formed between the hydroxyl group of HzpB28 and the carbonyl of Glu21' that we believe is responsible for the enhanced hexamer dissociation rates and delayed fibrillation behavior of HzpI. For example, an amino group in place of hydroxy at B28 should theoretically result in a stronger hydrogen bond with the carbonyl of Glu21' and if our hydrogen bond hypothesis is true, then NzpI may be an insulin variant that has a faster hexamer dissociation rate and is more stable than HzpI. Therefore, current work is focusing on screening libraries using parent synthetase mutant ProRS (M157Q) to improve Nzp incorporation. In parallel, screening for ProRS variants capable of improving Kep and Azp incorporation is also underway. Azp provides a chemical handle that allows for chemical conjugation via "click chemistry," which can further expand the repertoire for protein engineering through 'traceless' Staudinger ligation[58,59] reactions.

General method for screening aminoacyl-tRNA synthetases. Due to the length and amino acid sequence of β11, it can be postulated that the described screening system can be adapted for any cAA analog where the cAA is either not present in β11 or critical for its function and re-assembly with GFP(1-10). The necessary steps to adapt this system (for other projects involving new ncAAs) would be to clone the relevant synthetase in place of proS gene, and then proceed with library construction and subsequent screening.

Our goal for implementing a high-throughput screening system is to allow scientists to perform medicinal chemistry on proteins as a tool to understand protein behavior on an atom-by-atom level and eventually, to improve current or aid in the discovery of new therapeutics.

Materials and Methods

Materials. All canonical amino acids, (*2S,4S*)-4-amino-L-proline, 4-oxo-L-proline, and (*2S,4R*)-4-hydroxy-L-proline were purchased from Sigma. A portion of (*2S,4S*)-4-amino-L-proline was also purchased from Toronto Research Chemicals, Limited. (*2S,4S*)-4-hydroxy-L-proline was purchased from Bachem Americas and Pipecolic acid (Pip) was purchased from TCI America. All solutions and buffers were made using double-distilled water (ddH_2O).

Plasmid construction. All gblocks were ordered from Integrated DNA Technologies. Restriction enzymes, alkaline phosphatase, and ligase were purchased from New England Biolabs. All sequencing was submitted to Laragen for confirmation. All cloning-related transformation was done using either Mach1 (Invitrogen) or 10β cells (NEB). Miniprep plasmid scale and DNA purification was done using kits purchased from Zymo Research.

Large scale purification of plasmid DNA was done using Maxiprep kit purchased from Qiagen. Note that a silent mutation was made to proS to insert a *Sac*1 site (renamed proSs) via site-directed mutagenesis for ease of downstream cloning (library insertion).

Single plasmid. GFP(1-10) gblock with flanking Nhe1 sites and plasmid pQE80MCS-proSs was digested with Nhe1 for single-site digestion and ligation to produce plasmid pQE80MCS-proS_GFP(1-10). GFP11 gblock with flanking *BamH*1 and *Hind*3 sites and plasmid pQE80MCS-proS_GFP(1-10) were digested with *BamH*1 and *Hind*3 and purified. Linearized pQE80MCS-proS_GFP(1-10) was dephosphorylated and purified by gel extraction. Digested GFP11 and linearized vector was ligated to produce plasmid pQE80β11-proS_GFP(1-10).

Double plasmid. GFP(1-10) gblock with flanking BamH1 and Hind3 sites and plasmid pKYP680 (obtained from Tirrell Lab cell stocks; pBAD33 plasmid with entire Ara operon removed and used as an constitutive expression plasmid) was digested with BamH1 and Hind3 and purified. Linearized pKYP680 was dephosphorylated and purified by gel extraction. Digested GFP(1-10) gblock and linearized vector was ligated to produce plasmid p15a(J23100)_GFP(1-10) (herein onwards, pKYP680 is renamed p15a(J23100)_Wasabi). Promoters OXB20, *tac* and *uv5* were ordered as gblocks with flanking *Nsi1* and BamH1, and plasmids p15a(J23100)_GFP(1-10) and p15a(J23100)_Wasabi were digested with *Nsi*1 and *Bam*H1 and purified. Linearized p15a(J23100)_GFP(1-10) and p15a(J23100)_Wasabi were also dephosphorylated prior to purification to prevent re-ligation of the vector. Digested promoters and linearized vectors were ligated to produce the final constructs: p(*promoter)*-_GFP(1-10) and p(*promoter*)_Wasabi; all eight p15a constructs were confirmed by sequencing. All three variations of GFP11 constructs were purchased as gblocks with

flanking BamH1 and Hind3 sites and plasmid pQE80MCS-proSs was digested and ligated to produce pQE80(GFP11*)-proSs (WT, M157Q and C443G mutants included).

Library Construction. All libraries were cloned using electro-competent 10β cells.

Site-saturation mutagenesis. A fragment of the proS gene containing amino acids 120 to 200, with residues 157-161 randomized through NNK codons, was pieced together by sewing PCR using overlap ultramers. Vector backbone, pQE80β11(Eβ11)-proSs was purified from 100 mL culture by using Qiagen's Maxiprep Kit. The PCR library and vector backbone was digested with *Sac*1 and *Age*1. Linearized vector was dephosphorylated for three times longer reaction time to minimize recircularization (and avoid biasing library towards truncated products). Ligation reactions used 3 µg of digested backbone with a 3:1 vector-to-insert molar reaction. To enhance transformation efficiencies, we either purified ligation reactions using Zymo DNA concentrator spin columns or used ElectroLigase for compatibility for downstream electroporation. Cells transformed by electroporation were rescued in Super Optimal Broth with glucose (SOC) medium for 30 minutes at 37°C: a small volume was taken for plating and calculation of transformation efficiencies, and the rest was inoculated in rich medium (supplemented with ampicillin) overnight for downstream plasmid purification.

Error-prone. To avoid introducing mutations to the promoter and ribosomal binding site regions upstream of the start codon, the base pair just upstream of the start codon was mutated to introduce a *Sph*1 cut site (*proSs). Overhang primers containing *Sph1* and *Nhe1* sites were used to construct the error-prone library with the GeneMorph II Random Mutagenesis kit (Agilent Technologies). Error rate was adjusted by varying the amount of

starting template (pQE80β11(Eβ11)-*proSs (M157Q)) and quantified by sequencing. Two libraries were constructed (ep-1X and ep-3X) and error rate was determined by sequencing to be 9.8 ± 4.1 bp and 3.1± 1.3 bp mutants per proS gene, respectively.

A segmented error prone library was constructed using primers: BsrG1-R and Age1-F. Two libraries were constructed (LibA and LibB) and error rate was determined by sequencing to be 2.5 ± 1.5 bp and 1.5 ± 0.5 bp, respectively. Primers designed for error-prone PCR are as follows (corresponding the labels in Figure 5.8a):

- BsrG1-F: 5'-GATACCTTGTGCACGCAGTTCGCTGTACAGTT-3'
- Kpn1-F: 5'-TCGGAGTACTTGGTACCCAGCTGGAAGATGTG-3'
- Age1-F: 5'-GATAGAACCGGTGTCGGCTTGTACGGCGCGGA-3'
- Kpn1-R: 5'-GGTACCAAGTACTCCGAAGCACTGAAAGCCTC-3'
- Age1-R: 5'-TGGATTTCCGCGCCGTACAAGCCGACACCGGT-3'
- Sac1-R: 5'-TGATTCGTAACGAGCTCAGCTCTTACAAACAG-3'
- Sph1-R*: 5'-CTACCTTCCAACTGGAACCGTAACAGCATG-3'

Linearized vector was dephosphorylated for three times longer reaction time to minimize recircularization (and avoid biasing library towards truncated products). Ligation reactions used 3 μg of digested backbone with a 3:1 vector-to-insert molar reaction. To enhance transformation efficiencies, we either purified ligation reactions using Zymo DNA concentrator spin columns or used ElectroLigase for compatibility for downstream electroporation. Cells transformed by electroporation were rescued in Super Optimal Broth with glucose (SOC) medium for 30 minutes at 37°C: a small volume was taken for plating and calculation of transformation efficiencies, and the rest was inoculated in rich medium (supplemented with ampicillin) overnight for downstream plasmid purification.

Incorporation of proline analogs into GFP11. Proline-auxotrophic *E. coli* strain CAG18515 (CAG) was obtained from the Coli Genetic Stock Center at Yale University and made competent. CAG cells containing p15a plasmids were selected by chloramphenicol and further made competent to uptake pQE80 plasmids and grown on doubly-selective agar plate. Cultures containing either a single mutant or library of synthetases were grown overnight prior to 1:40 dilution into 1X M9, 20 aa medium (8.5 mM NaCl, 18.7 mM NH_4Cl, 22 mM KH_2PO_4, 47.8 mM Na_2HPO_4, 0.1 mM $CaCl_2$, 1 mM $MgSO_4$, 3 mg/L $FeSO_4$, 1 µg/L of trace metals (Cu^{2+}, Mn^{2+}, Zn^{2+}, MoO_4^{2-}), 35 mg/L thiamine hydrochloride, 10 mg/L biotin, 20 mM D-glucose, 200 mg/L ampicillin, 50mg/L chloramphenicol with 50 mg/L of L-amino acids, each). All expression experiments were conducted at 37°C, 200 RPM in shake flasks. At an appropriate cell density (OD_{600} ~0.8, growth phase of 3 hours), the culture was subjected to a medium shift; briefly, cells were centrifuged and washed with saline thrice prior to resuspension into 1.25X M9, 19 aa (1X M9, 20 aa medium without L-proline). After cells were further incubated for 30 min to deplete intracellular proline, 200 mL of 5X additives (1.5 M NaCl, 2.5 mM ncPro) was added to the culture. After another 15 min of incubation at 37°C to allow amino acid uptake prior to induction, IPTG was added to a final concentration of 1 mM. At the end of 1 h, cells were either read directly on a microplate reader or harvested by centrifugation for pre-processing prior to flow cytometry. Note that test co-expressions of GFP(1-10) and GFP11 followed the same protocol except no medium shift was performed, instead a 5X solution of NaCl and IPTG (1.5M and 5mM, respectively) was added to induce expression of GFP11.

Flow Cytometry. 1 h after IPTG induction of GFP11, cells were harvested by centrifugation and washed with 10X volume in sterile, cold PBS. Cells were strained using 0.4 µm nylon strainer to remove dead cells and other large particulates. FACS screening was performed on a MoFlo XDP cell sorter (Beckman Coulter, Brea, CA) using an argon laser (488 nm) and 530/40 nm bandpass filter. ProRS variants capable of incorporating ncPro were enriched by collecting the top 0.5% of cells characterized by the highest levels of fluorescence. Cells were sorted into SOC medium (1.5 mL), rescued at 37°C for 1 h, and further diluted into antibiotic-selective medium for overnight growth. A portion of the rescued, sorted cells was plated to obtain single colonies for sequencing purposes. Sorted cells not immediately used for further sorting were stored at -80°C in 25% glycerol or grown overnight for miniprep to obtain plasmid libraries. Naïve libraries were sorted using Purify1 mode and sorted cell populations were then subjected to a more stringent sort mode (Single).

References

1. Povolotskaya, I.S. & Kondrashov, F.A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465**(7300): 922-926 (2010).
2. Windle, C.L. et al. Extending enzyme molecular recognition with an expanded amino acid alphabet. *Proc. Natl. Acad. Sci. U. S. A.* (2017).
3. Fiacco, S.V. et al. Directed Evolution of Scanning Unnatural-Protease-Resistant (SUPR) Peptides for in Vivo Applications. *Chembiochem* **17**(17): 1643-51 (2016).
4. Roy, H. & Ibba, M. Molecular biology: Sticky end in protein synthesis. *Nature* **443**(7107): 41-42 (2006).
5. Haruna, K.-i., Alkazemi, M.H., Liu, Y., Söll, D. & Englert, M. Engineering the elongation factor Tu for efficient selenoprotein synthesis. *Nucleic Acids Research* **42**(15): 9976-9983 (2014).
6. Liu, C.C. & Schultz, P.G. Adding new chemistries to the genetic code. *Annu Rev. Biochem* **79**: 413-44 (2010).
7. Amiram, M. et al. Evolution of translation machinery in recoded bacteria enables multi-site incorporation of nonstandard amino acids. *Nat Biotech* **33**(12): 1272-1279 (2015).

8. Davis, L. & Chin, J.W. Designer proteins: applications of genetic code expansion in cell biology. *Nat Rev Mol Cell Biol* **13**(3): 168-82 (2012).
9. Yoo, T.H., Link, A.J. & Tirrell, D.A. Evolution of a fluorinated green fluorescent protein. *Proc. Natl. Acad. Sci. U. S. A.* **104**(35): 13887-90 (2007).
10. 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 Engl* **46**(28): 5340-3 (2007).
11. Truong, F., Yoo, T.H., Lampo, T.J. & Tirrell, D.A. Two-Strain, Cell-Selective Protein Labeling in Mixed Bacterial Cultures. *JACS* **134**(20): 8551-8556 (2012).
12. Wedemeyer, W.J., Welker, E. & Scheraga, H.A. Proline cis−trans isomerization and protein folding. *Biochemistry* **41**(50): 14637-14644 (2002).
13. Torbeev, V.Y. & Hilvert, D. Both the cis-trans equilibrium and isomerization dynamics of a single proline amide modulate β2-microglobulin amyloid assembly. *Proc. Natl. Acad. Sci. U. S. A.* **110**(50): 20051-20056 (2013).
14. Torbeev, V., Ebert, M.-O., Dolenc, J. & Hilvert, D. Substitution of proline32 by α-methylproline preorganizes β2-microglobulin for oligomerization but not for aggregation into amyloids. *JACS* (2015).
15. Lummis, S.C.R. et al. Cis-trans isomerization at a proline opens the pore of a neurotransmitter-gated ion channel. *Nature* **438**(7065): 248-252 (2005).
16. Lu, K.P., Finn, G., Lee, T.H. & Nicholson, L.K. Prolyl cis-trans isomerization as a molecular timer. *Nat Chem Biol* **3**(10): 619-629 (2007).
17. Kuemin, M. et al. Tuning the cis/trans Conformer Ratio of Xaa–Pro Amide Bonds by Intramolecular Hydrogen Bonds: The Effect on PPII Helix Stability. *Angew Chem Int Ed Engl* **49**(36): 6324-6327 (2010).
18. Erdmann, R.S. & Wennemers, H. Importance of Ring Puckering versus Interstrand Hydrogen Bonds for the Conformational Stability of Collagen. *Angew Chem Int Ed Engl* **50**(30): 6835-6838 (2011).
19. 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 Engl* **53**(39): 10340-10344 (2014).
20. Siebler, C., Trapp, N. & Wennemers, H. Crystal structure of (4S)-aminoproline: conformational insight into a pH-responsive proline derivative. *Journal of Peptide Science* **21**(3): 208-211 (2015).
21. Ibba, M. & Söll, D. Aminoacyl-tRNA synthesis. *Annu Rev. Biochem* **69**(1): 617-650 (2000).
22. Smith, T.F. & Hartman, H. The evolution of Class II Aminoacyl-tRNA synthetases and the first code. *FEBS Letters* **589**(23): 3499-3507 (2015).
23. Giulio, M. The evolution of aminoacyl-tRNA synthetases, the biosynthetic pathways of amino acids and the genetic code. *Origins of Life and Evolution of Biospheres* **22**(5): 309-319 (1992).
24. Delarue, M. Partition of aminoacyl-tRNA synthetases in two different structural classes dating back to early metabolism: Implications for the origin of the genetic

code and the nature of protein sequences. *Journal of Molecular Evolution* **41**(6): 703-711 (1995).
25. Fan, C., Ho, J.M.L., Chirathivat, N., Söll, D. & Wang, Y.-S. Exploring the Substrate Range of Wild-type Aminoacyl-tRNA Synthetases. *ChemBioChem* **15**(12): 1805-1809 (2014).
26. Tamura, K. Origins and Early Evolution of the tRNA Molecule. *Life* **5**(4): 1687-1699 (2015).
27. Stehlin, C., Heacock, D.H., Liu, H. & Musier-Forsyth, K. Chemical modification and site-directed mutagenesis of the single cysteine in motif 3 of class II *Escherichia coli* prolyl-tRNA synthetase†. *Biochemistry* **36**(10): 2932-2938 (1997).
28. Wong, F.-C., Beuning, P.J., Silvers, C. & Musier-Forsyth, K. An isolated class II aminoacyl-tRNA synthetase insertion domain is functional in amino acid editing. *J. Biol. Chem.* **278**(52): 52857-52864 (2003).
29. Johnson, J.M. et al. Multiple Pathways Promote Dynamical Coupling between Catalytic Domains in Escherichia coli Prolyl-tRNA Synthetase. *Biochemistry* **52**(25): 4399-4412 (2013).
30. Arnold, F.H., Wintrode, P.L., Miyazaki, K. & Gershenson, A. How enzymes adapt: lessons from directed evolution. *Trends in Biochemical Sciences* **26**(2): 100-106 (2001).
31. Chen, R. Enzyme engineering: rational redesign versus directed evolution. *Trends in Biotechnology* **19**(1): 13-14 (2001).
32. Swain, J.F. & Gierasch, L.M. The changing landscape of protein allostery. *Curr Opin Struct Biol* **16**(1): 102-108 (2006).
33. Raman, Arjun S., White, K.I. & Ranganathan, R. Origins of allostery and evolvability in proteins: A case study. *Cell* (166): 1-13 (2016).
34. Romero, P.A. & Arnold, F.H. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* **10**(12): 866-876 (2009).
35. Spiller, B., Gershenson, A., Arnold, F.H. & Stevens, R.C. A structural view of evolutionary divergence. *Proc. Natl. Acad. Sci. U. S. A.* **96**(22): 12305-12310 (1999).
36. Moore, J.C. & Arnold, F.H. Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. *Nat Biotech* **14**(4): 458-467 (1996).
37. Steiner, T. et al. Synthetic biology of proteins: Tuning GFPs folding and stability with fluoroproline. *PLoS ONE* **3**(2): e1680 (2008).
38. Kubyshkin, V. & Budisa, N. cis–trans-Amide isomerism of the 3,4-dehydroproline residue, the 'unpuckered' proline. *Beilstein Journal of Organic Chemistry* **12**: 589-593 (2016).
39. Kawahara-Kobayashi, A., Hitotsuyanagi, M., Amikura, K. & Kiga, D. Experimental evolution of a green fluorescent protein composed of 19 unique amino acids without tryptophan. *Origins of Life and Evolution of Biospheres* **44**(2): 75-86 (2014).
40. Wilson, C.G.M., Magliery, T.J. & Regan, L. Detecting protein-protein interactions with GFP-fragment reassembly. *Nat Methods* **1**(3): 255-262 (2004).
41. Cabantous, S., Terwilliger, T.C. & Waldo, G.S. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat Biotech* **23**(1): 102-107 (2005).

42. Cabantous, S. & Waldo, G.S. In vivo and in vitro protein solubility assays using split GFP. *Nat Methods* **3**(10): 845-54 (2006).
43. Toddo, S. et al. Application of split-green fluorescent protein for topology mapping membrane proteins in Escherichia coli. *Protein Sci.* **21**(10): 1571-1576 (2012).
44. Kamiyama, D. et al. Versatile protein tagging in cells with split fluorescent protein. *Nat Commun* **7**: 11046 (2016).
45. Nasu, Y. et al. Genetically encoded fluorescent probe for imaging apoptosis *in vivo* with spontaneous GFP complementation. *Analytical Chemistry* **88**(1): 838-844 (2016).
46. Ishikawa, H., Meng, F., Kondo, N., Iwamoto, A. & Matsuda, Z. Generation of a dual-functional split-reporter protein for monitoring membrane fusion using self-associating split GFP. *Protein Engineering Design and Selection* **25**(12): 813-820 (2012).
47. Shaner, N.C., Patterson, G.H. & Davidson, M.W. Advances in fluorescent protein technology. *J Cell Sci* **120**(Pt 24): 4247-60 (2007).
48. Zhu, Y. et al. Metabolic engineering of indole pyruvic acid biosynthesis in Escherichia coli with tdiD. *Microbial Cell Factories* **16**(1): 2 (2017).
49. 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. U. S. A.* **80**(1): 21-25 (1983).
50. Yuvienco, C., More, H.T., Haghpanah, J.S., Tu, R.S. & Montclare, J.K. Modulating supramolecular assemblies and mechanical properties of engineered protein materials by fluorinated amino acids. *Biomacromolecules* **13**(8): 2273-8 (2012).
51. Kowalczyk, T., Hnatuszko-Konka, K., Gerszberg, A. & Kononowicz, A.K. Elastin-like polypeptides as a promising family of genetically-engineered protein based polymers. *World Journal of Microbiology & Biotechnology* **30**(8): 2141-2152 (2014).
52. Roberts, S., Dzuricky, M. & Chilkoti, A. Elastin-like Polypeptides as Models of Intrinsically Disordered Proteins. *FEBS Letters* **589**(19 0 0): 2477-2486 (2015).
53. Mittelstaet, J., Konevega, A.L. & Rodnina, M.V. A kinetic safety gate controlling the delivery of unnatural amino acids to the ribosome. *JACS* **135**(45): 17031-17038 (2013).
54. Doerfel, L.K. et al. Entropic contribution of elongation factor P to proline positioning at the catalytic center of the ribosome. *JACS* **137**(40): 12997-13006 (2015).
55. Park, H.-S. et al. Expanding the Genetic Code of *Escherichia coli* with Phosphoserine. *Science* **333**(6046): 1151 (2011).
56. Aldag, C. et al. Rewiring translation for elongation factor Tu-dependent selenocysteine incorporation. *Angew Chem Int Ed Engl* **52**: 1441-1445 (2013).
57. Boon, K. et al. Isolation and functional analysis of histidine-tagged elongation factor Tu. *European Journal of Biochemistry* **210**(1): 177-183 (1992).
58. Saxon, E., Armstrong, J.I. & Bertozzi, C.R. A "Traceless" Staudinger Ligation for the Chemoselective Synthesis of Amide Bonds. *Organic Letters* **2**(14): 2141-2143 (2000).
59. Kiick, K.L., Saxon, E., Tirrell, D.A. & Bertozzi, C.R. Incorporation of azides into recombinant proteins for chemoselective modification by the Staudinger ligation. *Proc. Natl. Acad. Sci. U. S. A.* **99**(1): 19-24 (2002).

Acknowledgements

We thank M. Boyajian for collaboration in constructing ProRS mutants (notably, the M157 synthetase mutants and incorporation experiments), and development of the split-GFP screening system. We thank B. Silverman for flow cytometry assistance, discussion and collaboration in the development of the split-GFP screening system. We thank S. Lieblich for flow cytometry advice and discussion, and Y. Antebi for the EasyFlow software for flow cytometry analysis. We thank W. Glenn for editing this chapter.

Chapter 6 – Concluding Remarks

Proteins have naturally evolved from the 20 canonical amino acids (cAAs) with unique side-chains that participate in determining the structure of the macromolecule and therefore its function as well. The potential to create new proteins becomes near limitless with the introduction of non-canonical amino acids (ncAA) whose biorthogonal side chains can be synthetically-derived from all of chemical space.

Much of the focus in this thesis has centered on engineering insulin using ncAAs. Given the significance of diabetes within our society today, and predictions based on current trends suggesting cases of diabetes will grow even more substantially in the future, insulin is a highly relevant and important molecule to thoroughly understand and engineer. Compliance among diabetics is a known issue; developing insulins with a faster rate of onset is one avenue where there is an unmet need for the treatment of diabetes. We performed ncAA mutagenesis at a critical proline residue (but also a residue known to be uninvolved in insulin receptor binding and therefore, not critical for biological activity) and show that ncAAs can be applied to modulate the biophysical properties of a protein without adversely affecting its biological activity. Specifically, based on the side group, the biophysical properties of insulin can vary drastically. In the case of the hydroxy-insulins, we found that the kinetics of the hexamer dissociation and insulin stability can be improved; this was corroborated by structural studies finding evidence for a novel hydrogen bond within the insulin dimer interface. In the case of the fluoroinsulins, we find evidence of polar-π interactions that may be implicated in modulating insulin stability. Finally, by varying the ring-size of the proline surrogate ncAA, the properties of insulin can also be affected on a modular scale depending on the number of carbons in the prolyl ring at position B28.

Where limitations exist for protein engineering with cAAs, ncAAs may be able to offer a novel solution for which biology has never attempted to evolve. However, current methods for incorporation of ncAAs into newly synthesized proteins are limiting: not all ncAA can be utilized by a host organism's endogenous translational machinery, large-scale production may be difficult due to the practical and logistical demands of a media shift, and downstream processing (e.g. production of therapeutic insulin from *E. coli* requires proteolytic cleavage and chromatography purification) may result in a decrease in protein yield. The latter two will become more important engineering problems as use of ncAA gains in popularity within the scientific community and the demand of biologics increases (industrial-scale production to meet market demands). However, developing a system for expanding the repertoire of available ncAA possible to incorporate into proteins is of more current academic relevance because pilot experiments are generally on the laboratory benchtop scale. To this end, the work described in this thesis regarding a screening system for ProRS variants is of note. By expanding the possible proline-based amino acids, the insulin work described has potentially endless variants.

The application of ncAA mutagenesis to proteins is a general strategy. As our understanding of how to modulate protein properties and how to apply engineered therapeutics to combat diseases grows, the engineering strategies described in this work can be applied to a wider variety of targets or different problems on a known target. Already, there has been work on interferon-β, which is one of the major treatments for multiple sclerosis, where an azide-containing ncAA was introduced site-specifically to append a PEG chain to increase the drug's half-life *in vivo* and minimize the frequency of injections. Other work has focused

using the unique chemistry available to ncAAs in the design of potent antibody-drug conjugates that can specifically target cancer cells. In the case with insulin, there is great interest in the development of hepato-specific and glucose-responsive insulin technologies; natural evolution over the past millennia and recent experimental work limited to the cAAs has yet to yield any large advances; the ability to access biorthogonal chemistries may be necessary to achieve these unmet needs.

Although it is known that, in general, global replacement of a specific residue in a protein often has destabilizing consequences, there is a large body of literature demonstrating that laboratory evolution can be used to overcome these limitations. Therefore, it is reasonable to remain optimistic about ncAA mutagenesis as a tool for engineering therapeutic proteins. The success of engineering insulin with just several ncAAs is indicative that as protein engineering research continues to build, more tailored amino acids and effective methods for protein engineering will be discovered.