Chapter 4

FOLDTUNED PROTEINS ARE NOVEL AND FUNCTIONAL

4.1 Introduction

In the preceding chapter, we motivated and introduced "foldtuning" as a promising algorithm for generating far-from-nature and new-to-nature sequences that slot into the broad contours of natural fold families, leveraging structural plasticity in moderation as an extra, evolution-inspired source of structural and functional novelty. We also showed that, in addition to reflecting orthogonal-to-nature "language rules" for "writing" protein sequences, many foldtuned proteins pass a basic computational screen for physical reasonableness by exhibiting predicted-stable folded states.

In this chapter, we report on preliminary *experimental* validation for three foldtuned targets selected for amenability to high-throughput characterization, familiarity in the general field of protein science, and translational relevance for downstream synthetic biology and therapeutic applications. These three targets are as follows: (1) the SH3 domain, a small adaptor domain that mediates protein-protein interactions in receptor-initiated and cytoplasmic signal transduction pathways, often as part of tyrosine kinases (Kurochkina and Guha, 2013; Mayer, 2001); (2) the barstar fold, an antitoxin-like inhibitor of a secreted bacterial ribonuclease, the smallest and simplest of the known α/β folds, and additionally well-studied as a model system for concerted folding pathways and protein-protein interaction energetics (Schreiber and Fersht, 1995; Schreiber et al., 1994); and (3) insulin, the first peptide hormone discovered and characterized, the major regulator of anabolic metabolism in eukaryotes, and whose absence or dysregulation is the causative agent of diabetes (Mayer et al., 2007). Tailoring assays for expression, stability, and binding to the individualized circumstances of the aformentioned three fold targets, we demonstrate that foldtuned proteins are realizable and functional in certain *in vitro* and *in vivo* contexts. Augmenting these experiments with statistical and theoretical analyses of generated sequence architecture, structural features, and physicochemical properties we argue further that foldtuned models learn the minimal structural information required to maintain a core fold and to either (i) preserve existing function or (ii) broaden to novel ones depending on the selective pressure applied.

4.2 Results & Discussion

Foldtuned SH3 domains express stably

Emboldened by the ability of foldtuning to readily propose plausible far-from-natural protein sequences, we sought to validate selected examples experimentally for expression, stability, and function with minimal target-specific platform optimazation. From a roster of small folds ($\leq$ 84aa) for which coding DNA oligo pools could be easily synthesized, we focused first on the SH3-like barrel (SCOP ID: 2000090). The SH3 domain is a notable protein-protein interaction component and regulator of signal transduction, particularly in tyrosine kinase pathways. Engineered SH3 domains have historically been desirable in synthetic biology for roles in designed artificial protein recognition and signaling cascades, but attempts to develop an SH3 "toolkit" have been stymied by difficulties with *de novo* β-barrel design and off-target crosstalk with natural SH3s (Kim et al., 2023). SH3 structural homologs are strewn across functionally diverse superfamilies, including the aforementioned adaptor domains that commonly bind polyproline motifs in protein ligands, chromodomains that recognize histone methylated lysine marks, and large-subunit ribosomal proteins that scaffold rRNA.

Applying the standard evo+four foldtuning procedure to ProtGPT2 with SH3s as the target produced 2593 variants after *in silico* filtering, for a structural hit rate and sequence escape rate of 0.519 and 0.310 respectively. In contrast to, e.g. deep-mutational scanning libraries, proteins in foldtuned variant libraries — including for SH3s — boast high sequence diversity, featuring low pairwise sequence similarities and unique proteolytic digestion signatures (Fig. S4.1A-C). This enables direct high-throughput characterization of protein expression and select biophysical properties by mass-spectrometry-based proteomics without the additional complexity and cost of typical yeast-, mRNA-, or cDNA- display methods (Fig. 4.1A) (Rocklin et al., 2017; Tsuboyama et al., 2023). For our SH3 foldtuned library, 1347/2593 (51.9%) variants express at detectable levels in a reconstituted transcription-translation system as measured by untargeted mass-spectrometric profiling (Fig. 4.1B-C). Using length-normalized signal as a proxy for absolute abundance of expressed proteins, we observe signal intensity spanning $\sim$ 6 orders of magnitude, suggesting substantial variance in the intrinsic expressability of foldtuned SH3s and foldtuned designs more broadly; it must be emphasized, however, that these measurements cannot on their own account for confounding factors such as the imbalances in the makeup of the amplified oligo pool encoding the SH3 library. Regardless of this nuance, we see no evidence that expression level correlates with sequence similarity to natural

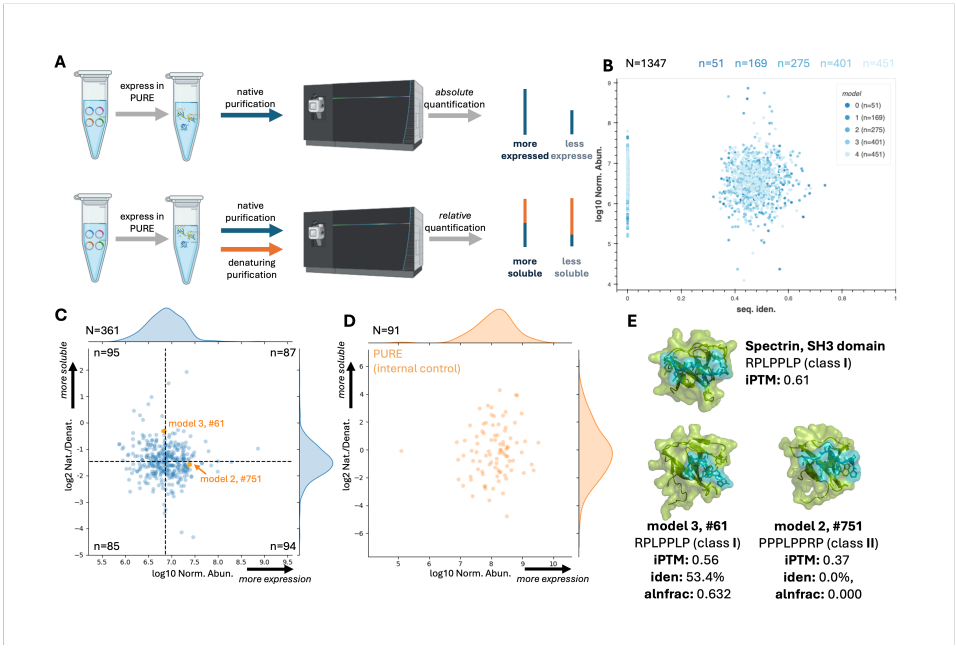

Figure 4.1: **Foldtuning-generated SH3s are expressable and stable.** (**A**). Schematic of mass-spectrometry-based proteomics assays for variant library expression and folding stability. (**B**) SH3 expression assay signal intensity normalized by expected tryptic peptide count vs. sequence identity to most-similar natural hit in UniRef50 for variants generated from models undergoing 0-4 rounds of foldtuning. (total $N = 1347$) (**C**) SH3 folding stability vs. expression assay results for $N = 361$variants detected in both contexts. Folding stability (y-axis) is measured by relative abundance ratio between natural and denaturing purification fractions. Normalized expression (x-axis) is measured as as in (**B**). (**D**) Folding stability vs. expression for PURExpress transcription-translation protein components ($N = 91$) as an internal control. (**E**) AlphaFold3 predicted structures and iPTM scores for selected SH3 variants (green) bound to a classI/II proline-rich peptide (teal), compared to the wildtype *G. gallus* spectrin SH3 domain.

SH3s or that it shifts based on the number of foldtuning cycles performed (Fig. 4.1B).

To rule out cases where high cell-free expression intensity might mask solubility and/or aggregation issues from poor folding stability we compared foldtuned protein recovery under native and denaturing purification conditions via multiplexed proteomics; variants without folding pathologies (e.g. exposed hydrophobic residues, buried and/or electrostatically clashing charged residues) are expected to show equivalent or greater signal in the native fraction relative to the denatured one (Fig. 4.1A). Analysis of the native/denatured signal fold-change for an internal control of $N = 91$ *E. coli* proteins originating from the reconstituted transcription-translation system

demonstrates that this stability/solubility proxy has a dynamic range spanning up to $\sim$ 10 orders of magnitude under the instrument conditions used (Fig. 4.1D).[1] Turning back to the foldtuned SH3s, 361 variants are detected confidently in both the absolute and multiplexed expression assay; absolute expression signal varies over $\sim$ 4 orders of magnitude, while the native/denatured signal fold-change varies over $\sim$ 6 orders of magnitude (Fig. 4.1C). Of particular interest is a subpopulation of 87 foldtuned SH3s that are both highly abundant in the initial expression assay and displaced away from the denatured fraction in the solubility/aggreggation assay, suggesting expressability, relative folding stability, and low aggregation propensity (Fig. 4.1C).

With an eye towards designing and optimizing SH3 parts for synthetic signal transduction, we computationally tested the hypothesis that foldtuned SH3 variants with high expressability and relative folding stability might recognize the proline-rich peptide motifs found in the full-length protein binding partners of natural SH3 domains (Mayer, 2001). *In silico* screening with AlphaFold3 predicts that, indeed, certain physically-plausible foldtuned SH3 variants can bind either class I or class II proline-rich ligands in a hydrophobic aromatic-sidechain-rich cleft analogous to the wild-type interface as exemplified by the *G. gallus* spectrin SH3 domain (Fig. 4.1E). Two exemplary foldtuned putative SH3s that emerge in the AlphaFold3-based screen are model 3 #61 (3_61) and model 2 # 751 (2_751). Variant 3_61 is a distant homolog of the guanine nucleotide exchange factor Vav (involved in cytoskeletal remodeling during lymphocyte development and activation) and is predicted to recognize the canonical class I motif RPLPPLP. Variant 2_751 has no detectable sequence homology to any known protein, yet is predicted to recognize the canonical class II motif PPPLPPRP.

To clarify how foldtuned models might be preserving critical structural and functional features in SH3s, including ones responsible for stability or binding of polyproline motifs, we turned to statistical coupling analysis (SCA). Originally developed to identify statistically interacting amino-acids from evolutionary-related sequence data, SCA has historically been applied to natural protein families to infer and extract physically connected "sectors" posited to comprise the minimal sequence information required to specify a fold and/or function (Halabi et al., 2009; Lock-

[1] The exact number and amino-acid sequences of protein components in the particular recombinant transcription-translation system used in this study, PURExpress from New England Biosystems, is proprietary. As a substitute approach, internal control proteins were mapped to an *E. coli* reference proteome.

less and Ranganathan, 1999; Socolich et al., 2005; Süel et al., 2003). Here, we applied SCA separately to natural SH3 domains and to the 2593 foldtuned putative SH3s and extracted sectors from each.[2] For both natural and synthetic SH3s, SCA finds a single small sector, covering eight and five residues in the natural and synthetic cases respectively (Fig. S4.2A-B). Natural and synthetic sectors are composed of non-overlapping sets of core residues; only a single sector position interacts directly with the bound proline-rich motif and it is shared between the natural and synthetic sectors. This suggests that, in line with the promiscuity and diversity of SH3-peptide binding, foldtuning may be preserving a bare-minimum sequence rule for binding few-among-many polyproline-like targets, while trying out a completely different solution for stably packing the SH3 β-barrel core. Ultimately, evaluating this interpretation will necessitate experimental validation of new foldtuning-enabled synthetic "links" in the SH3 connectome, potentially via a high-throughput/high-resolution SH3-peptide all-against-all cross-linking mass spectrometry approach.

Foldtuned barstars rescue bacteria from barnase toxicity

For a target with a more direct experimental readout of not just stability, but also function, we consider the barstar-like fold (SCOP ID: 2000624). With a single three-stranded parallel β-sheet packed against three α-helices, all connected by short loops, the barstar-like fold is an exceedingly simple α/β unit, familiar from foundational studies of protein folding stability (Schreiber and Fersht, 1995; Schreiber et al., 1994). The coding gene for its namesake protein, barstar, was originally identified in *Bacillus amyloliquefaciens* with orthologs distributed across gram-positive bacteria and structural homologs in the DNA double-strand break repair protein Mre11 and ribosomal protein L32e. Leveraging the expanse of the AlphaFoldDB, the custom SCOP-UniRef50 sequence-structure database also detects distant structural homolog barstar-like regions in proteins with putative ATPase and palmitoyltransferase activity, expanding the landscape of sequence motifs to harvest from. Barstar's native function in *B. amyloliquefaciens* is to inhibit, through a high-affinity active-site-occluding non-covalent interaction, the potent broad-spectrum bacterial ribonuclease barnase before its secretion into the surrounding environment.[3] Together, barnase (toxin) and barstar (antitoxin) comprise a toxin-antitoxin

[2] We believe this analysis and an analogous one on barstar later in the chapter to be the first examples of applying SCA to synthetic data to retrieve "pseudo"-evolutionary correlations.

[3] Barnase is so general a ribonuclease that its name is simply a portmanteau of "**ba**cterial" and "**r**ibo**n**ucle**ase**."

system, presenting an opportunity for functional screening of a foldtuned variant library for toxic gene rescue.

We apply the standard evo+four foldtuning approach to barstar, yielding 1403 variants after *in silico* filtering, for a structural hit rate and sequence escape rate of 0.281 and 0.560 respectively. Variants were co-expressed with barnase from *B. amyloliquefaciens* under a single *tac* promoter, under strong induction conditions, in a high *lacI E. coli* strain in order to mitigate confounding adaptations to barnase expression (Fig. 4.2A). In the absence of proper barstar expression and function, barnase expression is toxic to *E. coli* (Hartley, 2001). Functional foldtuned variants are expected to rescue host cells from the lethal effects of barnase expression. Comparing long-read sequencing counts of variant-coding amplicons, we found that 11 foldtuned barstar variants were significantly enriched ($p < 0.05$; Binyami-Hochberg correction for correlated tests) relative to uninduced (non-barnase-expressing) control under strong induction of barnase-barstar-variant co-expression, suggesting that the enriched variants are sufficiently functional mimics of barstar so as to mitigate the toxicity of barnase (Fig. 4.2B). Additionally, enrichment does not correlate with sequence identity relative to wild-type barstars or any natural protein. To this point, 7/11 of survival-enriched foldtuned barstars do not exhibit any detectable homology to natural sequences at the domain or sub-domain level (Fig. 4.2C).

For mechanistic insight and hypothesis refinement, we obtained AlphaFold3 predicted structures of the survival-enriched variants in complex with barnase. For four foldtuned variants — model 1 #633 (1_633), model 3 #647 (3_647), and model 4 #s 141 (4_141) and 219 (4_219) — these predicted complex structures indicate that barstar mimics are expected to bind barnase analogously to wild-type barstar, inserting an α-helix and adjoining loops into the binding pocket, obstructing the RNA hydrolsis active site (Fig. 4.2D). Detailed examination of predicted binding interfaces reveals that foldtuned barstars are expected to form hydrogen-bonds and salt-bridges with barnase, without steric or electrostatic clashes. Comparison with a published experimental structure of the endogeneous *B. amyloliquefaciens* barnase-barstar complex (pdb: 1BRS) suggests that fewer such contacts are expected with variants than with wild-type barstar, potentially indicating weaker binding and consequently reduced inhibition of barnase (Fig. 4.2D). It bears noting that this difference may stem at least in part from non-ideal bond geometries that persist due to AlphaFold3's lack of a side-chain or backbone relaxation step; molecular dynamics simulations could prove valuable for discriminating between binding strengths in a

more informative manner, as it has in a related design context for IL7RA minibinders (Lourenco et al., 2025).

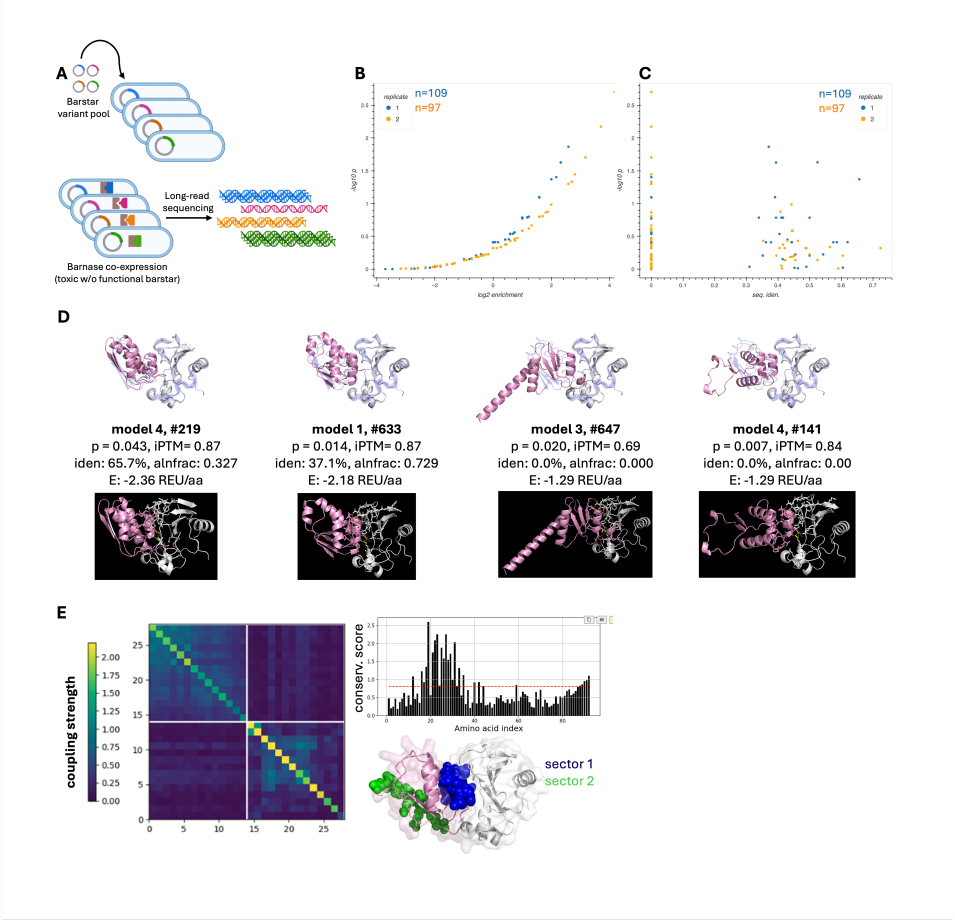

Figure 4.2: **Foldtuning-generated barstar variants are expressable, stable, and functional.** (**A**) Schematic of barnase-inhibition survival assay for barstar variant library stability and function. (**B**) Survival assay p-value rank plot for barstar variants. For a given variant, enrichment is calculated as the ratio of amplicon sequencing reads with and without induction of co-expression of the lethal binding-partner barnase (**C**). Survival assay p-values from (**F**) vs. barstar variant sequence identity to most-similar natural hit in UniRef50. (**D**) Top row: AlphaFold3 predicted structures, iPTM scores, and Rosetta energy predictions for selected barstar variants (pink) in complex with barnase (white). An experimental crystal structure of the wildtype barnase-barstar complex from *B. aquaforiensis* (pdb: 1BRS) is overlaid in blue. Bottom row: Predicted complex structures with putative hydrogen bonds and electrostatic interactions indicated. (**E**) Results of statistical coupling analysis (SCA) on $n = 1493$ foldtuned barstar sequences. Left: Second-order coupling matrix, blocked into two orthogonally co-evolving sectors. Top right: First-order conservation scores. Bottom right: Visualization of sector positions mapped onto a representative barstar-barnase complex structure (pdb:2ZA4).

Looking beyond the interface, AlphaFold3 predictions also imply that some variants may explore alternate conformations and binding modes in the neighborhood of expected barstar structure and function, exemplified by 4_141, a zero-homology variant that is predicted to have low β-content and rotate 90° relative to the wild-type while maintaining α-helical and loop elements in the barnase binding pocket (Fig. 4.2D). Of note, the two variants predicted to adopt more dramatically different conformations — 4_141 and 3_647 — are also assigned Rosetta energy scores on the higher end of plausible; in the absence of experimental structure determination, these structure-function hypotheses should be taken with a degree of caution.

Given the detection of foldtuned barstar mimics with antitoxin-like function, and circumstantial indications that at least some among these mimics may utilize similar structural solutions to wild-type barstar, a natural question to ask is that of what sequence and/or structure "rules" the foldtuned models themselves have learned? Multiple sequence alignment of wild-type barstar along with the eleven survival-enriched foldtuned variants reveals that in the contiguous nineteen-residue region (columns 38-56) spanning the barnase-binding interface, toxicity-rescuing variants preserve 6-11 (32-58%) of wild-type amino-acid identities (Fig. S4.3). Clearly, foldtuned models are not simply memorizing the semantics of barnase-binding and scaffolding them into redesigned flanks.

For a deeper view of how foldtuned models might be preserving the structural-functional "grammar" of barstar, we return to SCA, this time treating the 1403 foldtuned barstar variants as a synthetic protein family.[4] SCA proposes two sectors; one, at the C-terminus, is most likely an artifact attributable to ESM batch-inference token-padding with residual alanines; the other maps onto the barnase-binding interface (Fig. 4.2E). This suggests that foldtuning has "solved" the barnase-binding problem by decoupling the critical inserted α-helix motif from the rest of the protein, preserving only its most salient sequence features, and inventing wholly new ways to fill in the remainder of the barstar fold. In other words, foldtuning has distilled the structural and functional nature of barstar into a single essential grammar rule.

Foldtuned and PLM-sampled insulins are INSR binders and agonists

Lastly, we steered foldtuning to design mimics of insulin, a high-value translational target well outside of our initial set of 727 SCOP folds, posing several new

[4]Unlike in the SH3 case, we do not conduct the comparative analysis on natural barstars, as the wild-type sequences are so highly conserved across gram-positive bacteria as to swamp out the second-order interaction signatures that SCA relies on.

challenges for the foldtuning algorithm and workflow to overcome. Insulin hardly requires introduction as a protein of interest. It is the preeminent peptide hormone in all eukaryotes; insulin signaling coordinates anabolic metabolism across cells, tissues, and organs (Mayer et al., 2007). Its absence (due to self-reactive destruction of most insulin-producing pancreatic β-cells) is the salient aspect of type I diabetes; its dysregulation (insulin resistance) underlies type II diabetes. Insulin presents multiple challenges to the thematic underpinnings and practical application of foldtuning. To the former aspect, the active form of insulin is deeply conserved across eukaryotes at the sequence level, and shares a structural neighborhood with related peptide hormones including insulin-like-growth-factor-1 (IGF-1), relaxins, and several insulin-like peptides (ILPs) of unclear function; insulin and IGF-1 preferentially bind to and are agonists of their cognate receptor tyrosine kinases, with weak cross-reactivity; relaxins and ILPs cross-react with several GPCRs (Claeys et al., 2002). This suggests that foldtuning, with its emphasis on innovation about a template structure in moderation, may sample a range of specific and promiscuous binding phenotypes as opposed to binders specific to the insulin receptor INSR.

As far as practical implementation obstacles, insulin is a tricky target for foldtuning thanks to the post-translational internal cleavage events required to transform inactive, largely disordered proinsulin into structured, active insulin through excision of the C-peptide (which makes up 31 of the 86 residues in the coding region of the INS gene) and formation of three disulfide bonds (two interchain between the A- and B-peptides; one intrachain within the A-peptide). To circumvent this issue and to align with standard expression and characterization processes in industry, we foldtunded ProtGPT2 to generate single-chain insulin variants that are fusions of the A- and B-peptides. Natural training sequences ($n = 335$, reduced to $n = 193$ after deduplication clustering) and reference structure fragments were taken from InterPro entry IPR004825, which ostensibly includes insulin and excludes IGF-1, relaxin, and ILPs, though sequence homology considerations cast some doubt on the robustness of this filtering. IPR004825 sequences were multiply aligned to *H. sapiens* insulin to identify putative C-peptide regions to be removed before clustering and downsampling, leaving single-chain A/B fusion training data. Standard evo+four foldtuning rounds yielded 2889 putative insulin variants with structure hit and sequence escape rates of 0.578 and 7×10^{-4} respectively. The atypically low sequence escape hit for foldtuned insulin models (only 2/2889 variants lacking detectable homology to natural proteins), as well as a median 80.0% sequence similarity to the closest natural hit, likely stems from the aforementioned high degree

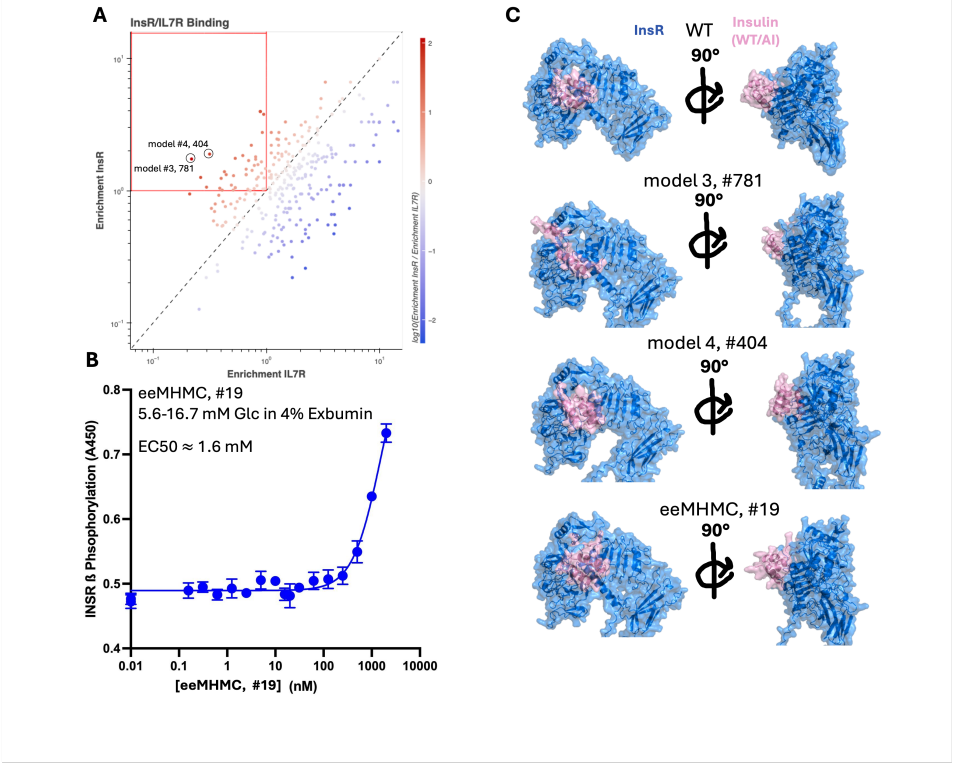

Figure 4.3: **PLM-designed insulin variants bind the endogenous insulin receptor.** (**A**) Relative enrichment plot of foldtuned insulin variant binding to the endogenous INSR receptor (on-target) vs the endogenous IL7R receptor alpha-chain (off-target) using the Protein CREATE platform. (**B**) Sandwich ELISA (A450 spectrophotometric readout) of INSR phosphorylation in response to explore-exploit Metropolis-Hastings Monte Carlo (eeMHMC) designed insulin variant #19 (estimated EC50=1.6 mM). (**C**) AlphaFold3 predicted structures of WT *H. sapiens* insulin and experimentally plausible designed INSR binders (pink) in complex with the native INSR receptor ectodomain (blue).

of sequence conservation among natural insulins and from a tradeoff in choosing a small InterPro family from which to initiate foldtuning in hopes of prioritizing INSR-specific binders and agonists.

We used the Protein CREATE platform to screen all foldtuned putative insulins for INSR-specific binding as described in Lourenco et al. (2025). In brief, variants are displayed on T7 bacteriophage and screened against multiple receptor candidates ligated to magnetic beads, with a sequencing-based readout of amplicon counts before- and after- receptor-bead pulldown, resulting in a vector of enrichment scores for each receptor screened. Here we screen against two receptors, taking enrichment

after INSR pulldown as a measure of on-target binding, and enrichment after IL7RA (a type I cytokine receptor) pulldown as a measure of generic off-target binding. For 307 foldtuned variants with sufficient reads to calculate enrichment in both contexts, 41 (13.4% of detected; 1.4% of entire foldtuned pool) are enriched (enrichment > 1) for INSR-binding and de-enriched (enrichment < 1) for IL7RA-binding (Fig. 4.3A). This is fewer than the 61 variants (19.9% of detected; 2.1% of entire pool) with the inverse phenotype of IL7RA-binding enrichment and INSR-binding de-enrichment. The lion's share of variants — 159 (51.9% of detected; 5.5% of entire pool) exhibit a doubly-enriched phenotype in this assay — underscoring, altogether, the inherent difficulty of designing *specific* binders, as well as an opportunity in applying foldtuning as a zero-shot generator of binding *phenotype diversity* suitable for multiple downstream optimizations.

To validate INSR-binding activity, we attempted chemical synthesis of the two foldtuned variants with the highest relative enrichment scores (INSR enrichment / IL7RA enrichment) out of the 41 variants that display the INSR-specific phenotype in Protein CREATE screening — these are model 3, #781 (3_781) and model 4, # 404 (4_404). However, neither variant is observed to refold solubly following synthesis and denaturation, suggesting a failure of proper disulfide bond formation, pointing potentially to a lack of a uniquely stable ground state among multiple cyclization isomers; troubleshooting is ongoing. AlphaFold3 predicts that both priority variants should indeed bind to the INSR ectodomain with ligand-receptor contacts reminiscent of but not identical to those formed by wild-type insulin, lending support to the emerging paradigm from our investigations of SH3 and barstar that foldtuning retains only those sequence rules minimally necessary for marginal binding while injecting novelty that percolates to perturbed contacts, pockets, and downstream phenotypes (Fig. 4.3C).

The challenges presented for foldtuning by insulin's sequence and structure features led us to develop and evaluate a second PLM-based generation strategy in parallel. We call this strategy "**e**xplore-**e**xploit **M**etropolis-**Ha**stings **M**onte **C**arlo", or eeMHMC. We are far from the first to take an MHMC or Markov Chain Monte Carlo (MCMC) approach more generally to the problem of sampling over a sequence landscape from an encoder-only PLM. What is novel about our approach is the energy function that determines the acceptance probability of proposed moves. Where others have written energy functions that consider only the likelihood of a given sequence as inferred by a PLM, sometimes regularized to explicitly favor memo-

rization of natural sequence motifs, we propose a two-term function that balances leveraging what the PLM has internalized about sequence plausibility (the "exploit" term) with an incentive to make large semantic changes (the "explore" term) (Hie et al., 2022; Verkuil et al., 2022).[5] In this way, eeMHMC stands on common ground with the grammar-respecting/semantics-altering behavior of foldtuning. We generate 100 candidate insulin mimics with ESM2-650M-based eeMHMC, seeded by column-wise independent sampling from a deep insulin MSA. Among these 100 variants, sequence identity to human insulin drops as low as 31.4% while still respecting the canonical insulin fold (including disulfide staples) and showing reasonable surface hydrophobicity and electrostatics, according to computational analysis on predicted structures (Fig. S4.4). Outside collaborators attempted to express 20 of the eeMHMC variants under standard industrial conditions; of these, one variant, eeMHMC_19 (56.8% identical to human insulin), was able to be expressed, purified, and refolded. Furthermore, eeMHMC_19 shows agonist activity for INSR according to sandwich ELISA readout of receptor phosphorylation (Fig. 4.3B). Although this activity is weak, with an inferred EC50 of 1.6 mM, $\sim$ 1000x higher than for wild-type insulin and $\sim$ 10-50x higher than for wild-type IGF-1, it still represents remarkable progress towards a functional insulin mimetic with substantially reduced resemblance to wild-type.

4.3 Conclusion

Picking up from *in silico* revelation of the potential scope and breath of foldtuning as a synthetic biology design engine, we completed a fastidious series of experiments, supported by mechanistic insight from computation, underscoring that foldtuned proteins are novel, realizable, and functional. Taking the SH3 domain as an initial representative and experimentally tractable fold, we showed that foldtuning clears initial but essential criteria of generating expressable, stable, non-aggregation-prone proteins. Raising the bar to a toxin-antitoxin system where structure and function are closely linked, and where selective pressure disfavors natural exploration, we found that foldtuning maintains function as an apparent byproduct of preserving structural constraints, recapitulating the fold-and-function "grammar" of the barnase-binding interface of barstar with novel semantics. Coming to insulin, an ambitious target with a host of obstacles that could have been erected deliberately to thwart foldtuning — sequence deeply and anciently conserved, structure shared across a bucket of related peptide hormones, significant processing required *in vivo* to go from an

[5] Full mathematical details are reported in Section 4.4.

inactive single-chain propeptide to an active multi-chain species — we still obtained INSR-receptor-specific binder candidates. Moreover, we introduced a second PLM-based generation mechanism, explore-exploit MHMC, which, like foldtuning, perturbs sequence semantics in a stepwise fashion while implicitly retaining structural grammar, and applied it to design and validate a novel INSR agonist.

Beyond their utility as interesting and convenient case studies, the three targets examined in detail here proffer an assortment of pertinent takeaways for the future of foldtuning. The apparent allowance of foldtuning for recognizing variable sequence motifs on binding-partners implies that pools of foldtuned SH3s could contain enough diversity to mine, optimize, and organize sets of orthogonal nature-inspired parts with new-to-nature recognition logic into synthetic signaling cascades or larger fully synthetic connectomes. Such an approach would port over conveniently to other highly modular binding domains such as SH2s and PDZs. Similarly, the insulin-binding results are a powerful reminder of the many degrees of freedom in (ant)agonist design — an information-rich way forward will be to embrace the "novelty first, fitness next" mindset of foldtuning and screen putative binders against whole receptor repertoires to discover brand-new phenotypes in cell signaling space.

4.4 Methods

Oligo Pool Design and Preparation

Foldtuning-generated sequences selected for experimental characterization were truncated to remove disorded N- and C-terminal tail regions as predicted by ESMFold and identified in C_α contact maps computed with BIOTITE. Coding DNA sequences were designed by reverse translation with DNACHISEL, codon-optimizing for *E. coli*, with additional constraints on GC content (global ≥ 0.25, ≤ 0.65; never ≤ 0.19 or ≥ 0.71 over any subsequence of length 50) and homopolymers (restricted to < 14nt). Constant flanks — GACTACAAGGACGACGATGACAAG (5') and GGTTCCCACCATCATCACCATCAT (3') were added to code for a 5' FLAG tag and a 3' GSHHHHHH tag.

Oligo pools were ordered from Twist Biosciences as ssDNA fragments for sequences $\leq$ 300nt or as dsDNA fragments for sequences $>$ 300bp and PCR-amplified with Q5 Hot Start High-Fidelity 2X Master Mix (NEB, M0494S) according to manufacturer instructions. T7RNAP promoter, ribosome binding site, start codon, stop codon, and T7 terminator elements were added in a subsequent PCR-amplification step with the same reagents, and purified, concentrated, and resuspended in ultra-pure water

using the Monarch Spin PCR & DNA Cleanup Kit (NEB, T1130S) according to manufacturer instructions.

In vitro Expression Measurements

Foldtuned variant pools were expressed *in vitro* with PURExpress (NEB, E6800) following the manufacturer's protocol, with 500ng template dsDNA per 50 µL reaction volume, incubating 18hrs at 29 °C. Expressed protein was purified under native conditions by His-tag pulldown using NEBExpress Ni Spin Columns (NEB, S1427L); 400 µL of eluate was washed and concentrated with Amicon Ultra Centrifugal Filters, 3 kDa MWCO (Millipore, UFC5003) 4x with 400 µL phosphate-buffered saline pH7.4, centrifuging at 14,000*g* for 30min per exchange, and 50 µL of concentrate recovered by reverse spin (1000*g* for 2min).

Concentrated purified protein samples were digested in an S-Trap micro spin column (Protifi, USA) according to the manufacturer's instructions and analyzed on Q-Exactive HF mass spectrometer coupled to EASY-nLC 1200. Peptides were separated on an Aurora UHPLC Column (25 cm × 75 µm, 1.7 µm C18, AUR3-25075C18-TS, Ion Opticks) with a flow rate of 0.35 µL/min for a total duration of 1hr and ionized at 2.2 kV in the positive ion mode. Raw data files were searched against the Uniprot Escherichia coli proteome (UP000531813) and foldtuned variant sequences. Searches used the Proteome Discoverer 2.5 software based on the Sequest HT algorithm. Oxidation / +15.995 Da (M), deamidation / +0.984 Da (N), and acetylation / +42.011 Da(N-term) were set as dynamic modifications; carbamidomethylation / +57.021 Da (C) was set as fixed modification. The precursor mass tolerance was set to 10 ppm, whereas fragment mass tolerance was set to 0.05 Da. The maximum false peptide discovery rate was specified as 0.01 using the Percolator Node validated by q-value. Absolute abundance signal intensities were scaled by dividing by the expected peptide count from simulated tryptic digestion.

In vitro Folding Stability Measurements

Foldtuned variant pools were expressed, purified, washed, and concentrated as for the expression assay, as described above, with the modification that the reaction volume was split post-expression into 2×25 µL aliquots, one purified under native conditions and the other under denaturing conditions (6 M guanidinium chloride) following manufacturer instructions.

Concentrated purified protein samples were analyzed by Eclipse mass spectrometer coupled to Vanquish Neo. 1ug of peptides from S-trap based digestion with TPCK-

treated trypsin were injected and separated on an Aurora UHPLC Column (25 cm × 75 μm, 1.7 μm C18, AUR3-25075C18-TS, Ion Opticks) with a flow rate of 0.35 μL/min for a total duration of 1 hour and ionized at 1.8 kV in the positive ion mode. Raw data files were searched against the Escherichia coli (strain B / BL21-DE3) proteome (UP000002032) foldtuned variant sequences using the Proteome Discoverer(PD) 2.5 software based on the SequestHT algorithm. Oxidation / +15.995 Da (M), Deamidated / +0.984 Da (N, Q), acetylation / +42.011 Da (protein N-term) and Met-loss / -131.040 Da (protein N-term, M) were set as dynamic modifications, and carbamidomethylation / +57.021 Da (C) was fixed modification. The precursor mass tolerance was set to 10 ppm, whereas fragment mass tolerance was set to 0.6 Da. The maximum false peptide discovery rate was specified as 0.01 using the Percolator Node validated by q-value. Enrichment was calculated as the abundance ratio of the natural channel relative to the denatured channel.

Barstar-Barnase Survival Assay

The barstar-like foldtuned variant pool was designed, ordered, and amplified to add regulatory elements as described above. Barstar variants were cloned as a single pool into barnase-barstar expression vector pMT416 (gift from Robert Hartley, Addgene plasmid #8607; http://n2t.net/addgene:8607; RRID:Addgene_8607), replacing the wild-type barstar-coding region, using NEBuilder HiFi DNA Assembly Master Mix (NEB, E2621S) according to manufacturer's instructions. 1 μL of assembly product was transformed into 10 μL 5-alpha F'Iq Competent *E. coli* (NEB, C2992I) following the standard manufacturer heat-shock protocol. Outgrowth product was used to seed 2mL LB cultures at 1-in-200 dilution and incubated overnight at 37 °C, 250 rpm with carbenicillin as the selection marker. Upon reaching an OD600 of 0.6, cultures were split into two 1 mL aliquots; 1mM IPTG was added to one aliquot per pair, the other was kept as an untreated control; all aliquots were incubated at 37 °C for 3hrs to strongly induce protein expression. Barstar-variant-coding regions were amplified directly from 0.2 μL of culture using Q5 Hot Start High-Fidelity 2X Master Mix (NEB, M0494S). PCR product was purified as described above, diluted to 5 ng/μL, and Premium PCR Sequencing performed by Plasmidsaurus using Oxford Nanopore Technology with custom analysis and annotation.

Reads were translated and filtered to retain only protein sequences containing the expected N- and C-terminal tag leader sequences and not prematurely truncated by a misplaced STOP codon. Translated reads were mapped back to the foldtuning-generating barstar variant sequences with MMSEQS2, requiring an aligned region

of > 80aa with a minimum sequence identity of 98%. Variant enrichment was calculated as the ratio of mapped reads under barnase-barstar induction vs the uninduced control. P-values were computed non-parametrically by assuming a null model of random read allocation, drawing 10^6 samples.

Bioinformatics Analysis

Multiple sequence alignments (MSAs) were calculated using MUSCLE v5 via the EMBL-EBI webserver (Edgar, 2022).

Statistical coupling analysis (SCA) was performed with pySCA v6.1 and visualizations created with PyMOL v3.1.0 (Rivoire et al., 2016).

Energy Scoring Calculations

Biomolecule energy scores were obtained using the default 'ref2015' energy function and standard relaxation and scoring workflow in Rosetta v3.11, as described in Alford et al. (2017). Energy scores are reported in **R**osetta **E**nergy **U**nits (R.E.U.), normalized to sequence length.

Binding Mode Prediction and Analysis

Unless specified to the contrary, AlphaFold3 was used for all structure prediction tasks involving protein-protein or protein-peptide complexes, via the AlphaFold-Server interface (https://alphafoldserver.com). For the SH3 domain, predicted complex structures were computed for foldtuning-generated putative SH3 variants in the presence of a representative class I (RPLPPLP) or class II (PPPLPPRP) proline-rich peptide motif. For the barstar-like fold, predicted complex structures were computed for foldtuning-generated putative barstar variants in the presence of wild-type barnase from *B. amyloliquefaciens*(uniprot:P00648). Predicted structures were compared to a wild-type reference, either the spectrin SH3 domain from *Gallus gallus* or the barnase-barstar complex from *Bacillus amyloliquefaciens* (pdb: 1brs). For insulin, predicted complex structures were computed for foldtuning-generated and/or PLM-sampled putative insulin variants in complex with the monomeric full-length ectodomain of human INSR (insulin receptor).

All predicted structures were visualized with PyMOL v3.1.0. For the barnase-barstar complex, good hydrogen-bonds, acceptable hydrogen-bonds, and electrostatic clashes were inferred and displayed with the PyMOL "show_contacts" third-party plugin. For insulins, hydrophobicity was visualized using the "color_h" third-party plugin and electrostatic potential was calculated and visualized using the

APBS Electrostatics plugin.

High-Throughput Insulin Binding Assay

A library of 2889 insulin variant amino-acid sequences was constructed by foldtuning on InterPro entry IPR004825, containing 335 natural insulin sequences (reduced to 193 sequences after deduplication clustering at 100% similarity with MMSEQS2) integrated from overlapping entries in the PRINTS, CDD, and PANTHER databases. Foldtuning was executed as described in Section 3.4, with the modification that generated variants were post-processed by aligning to the sequence *H. sapiens* insulin (uniprot: P01308) and removing residues aligning to the C-peptide region that is removed by proteolytic cleavage *in vivo* during the conversion of inactive proinsulin to active insulin, resulting in a library of *single-chain* insulin mimics.

High throughput binding measurements (sequencing read enrichment scores) were obtained using the Protein CREATE platform as described in Lourenco et al. (2025) with INSR as the on-target receptor and IL7RA as the off-target decoy receptor.

Insulin Variant Generation by MHMC Sampling

Additional insulin variants (not screened with Protein CREATE) were generated through Metropolis-Hastings Monte Carlo (MHMC) sampling from an insulin-like sequence landscape with an two-term energy function combining a preference for accepting mutations that increase sequence-likelihood under the ESM2-650M model (the "exploit" term) with a preference for accepting mutations resulting in a large semantic change relative to the current sequence (the "explore" term). An individual sequence s_i of length N has an associated log-likelihood $L_i = \prod_{k=1}^{N} l_k$ where the l_k represent indepedent residue-wise likelihoods, and an ESM2-650M final-layer mean-pooled embedding vector $\mathbf{x_i}$. Semantic change is defined as $S_{i\to j} = \|\mathbf{x_j} - \mathbf{x_i}\|_1$ for a pair of sequences s_i, s_j. As semantic change is not defined for individual sequences, it is not possible to define an absolute sequence energy E_i; it is however possible to define $\Delta E_{i\to j}$ for a proposed move from s_i to s_j. Precisely

$$\Delta E_{i\to j} = (\log L_j - \log L_i) + w_s S_{i\to j} \tag{4.1}$$

where $L_i, L_j, S_{i\to j}$ are defined as above and w_s is a coefficient that controls the relative weights assigned to the exploit and explore terms.

The standard Metropolis-Hastings acceptance criterion is used; namely, a proposed move (restricted in this method to a single point mutation, sampled uniformly

across sequence positions and amino-acid identities) from s_i to s_j is accepted with probability

$$p_{i \to j} = \min\{1, \exp(\beta \Delta E_{i \to j})\} \tag{4.2}$$

where β refers to the thermodynamic β, the inverse of the sampling temperature T.

In all variant generation runs for this study, MHMC was run for $n = 2000$ steps, $w_s = 0.4$, and adaptive temperature adjustment at 100-step intervals. For each run, the intial sequence s_0 was sampled column-wise from a multiple sequence alignment of 687 wild-type insulin sequences obtained by querying UniProt for all matches to the INS gene, with individual amino-acid sampling probabilities proportional to the amino-acid distribution over each individual column. Sampled gap characters and putative C-peptide regions were removed prior to concatenation into s_0.

A total of 100 independent MHMC replicates were performed. Sequences with fewer than 4 or an odd number of cysteine residues were removed. Sequences were converted to single-chain fusions by inserting a GGGRGG loop in-between the concatenated A-peptide and B-peptide The resulting sequences were binned into four quadrants by sequence identity % to *H. sapiens* insulin and predicted iPTM score in complex with the INSR receptor ectodomain according to AlphaFold-Multimer prediction, with 20 variants across the quadrants forwarded for attempted expression, refolding, and activity characterization.

Expression and Validation of Insulin Mimic Cellular Activity

Twenty insulin mimics designed via "explore-exploit" Metropolis-Hastings Monte Carlo (eeMHMC) sampling from the ESM2-650M model were expressed in *E. coli* and refolding was attempted as previously reported in the literature (Chen et al., 2016; Min et al., 2011). Refolded monomers were isolated by reverse-phase high pressure liquid chromatography (RP-HPLC) and evaluated for insulin receptor agonist activity using a sandwich ELISA INSR-β subunit assay specific for receptor tyrosine residue phosphorylation (Maloney et al., 2003). EC50 values were inferred from a four-parameter logistic regression model fit to 450 nm absorbance vs. variant concentration data.

4.5 Supplemental Material

Supplemental Figures

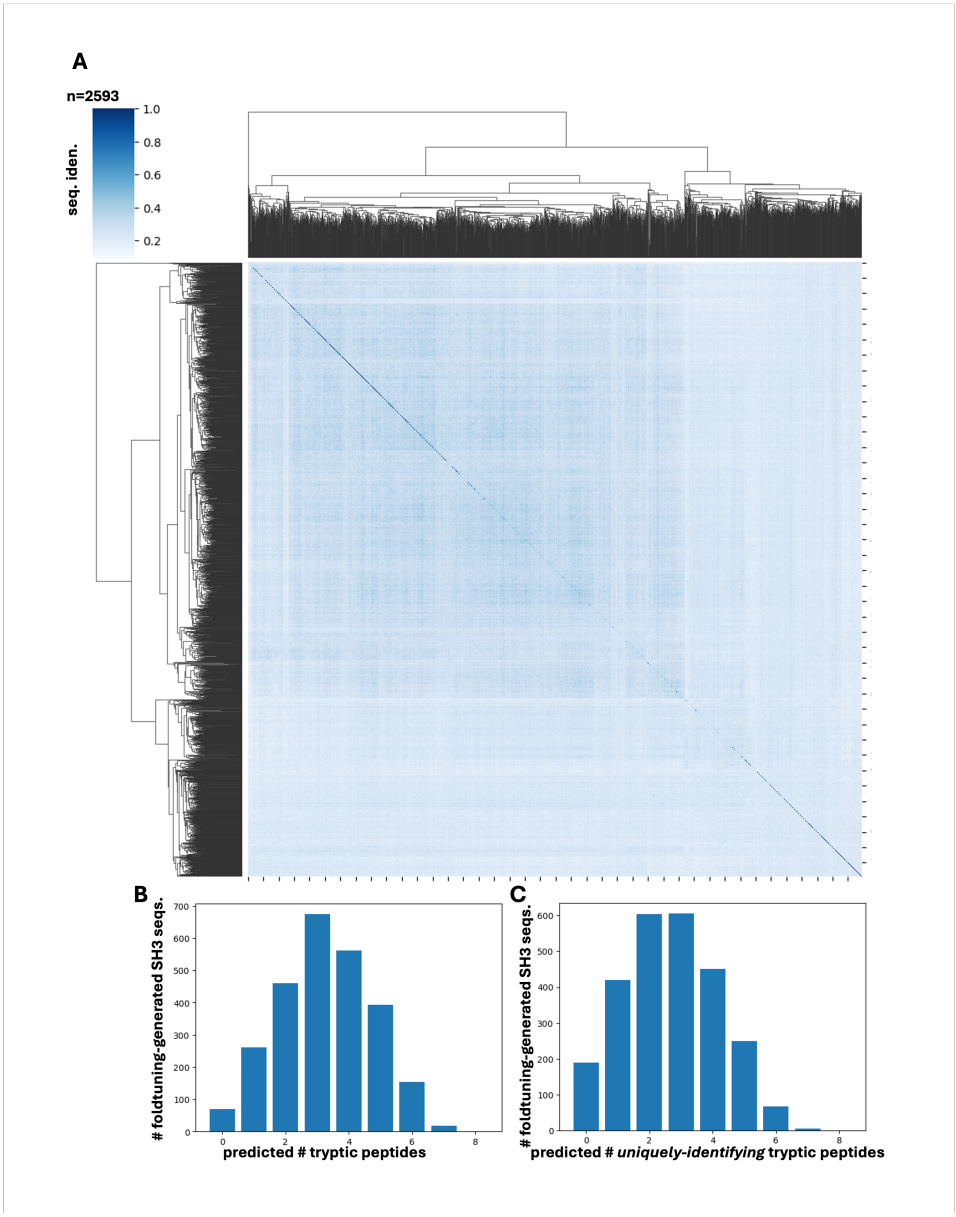

Figure S4.1: **Sequence diversity and detectability of foldtuning-generated SH3 domains.** (**A**) Hierarchically clustered heatmap of pairwise sequence identity between $n = 2593$ SH3 domain candidate sequences generated via foldtuning. (**B**) Expected detectable peptide counts predicted by in silico tryptic digestion. (**C**) Counts of predicted tryptic peptides that map uniquely to single foldtuned SH3 variants.

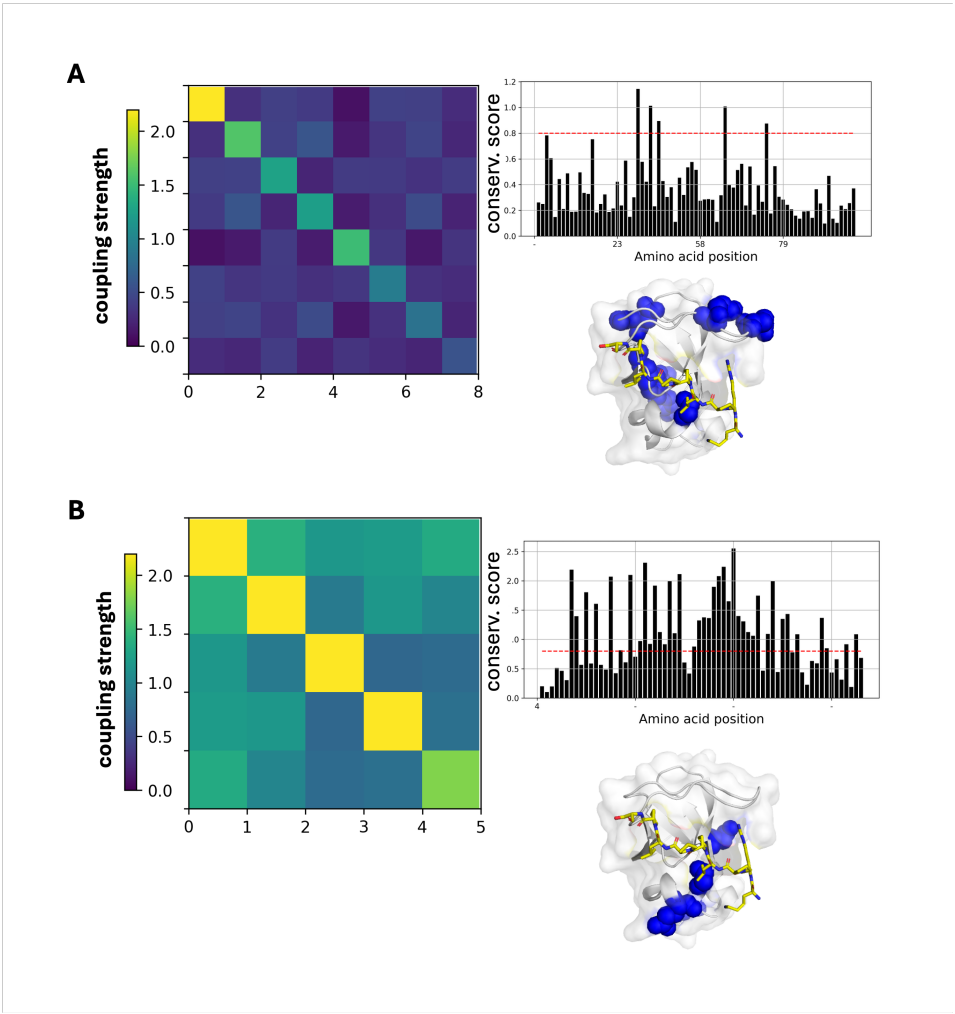

Figure S4.2: **Statistical coupling analysis of natural and synthetic SH3s.** (**A**) Results of statistical coupling analysis (SCA) on $n \approx 2500$ natural SH3 domain sequences. Left: Second-order compressed coupling matrix, blocked into a single statistically interacting sector. Top right: First-order conservation scores. Bottom right: Visualization of sector positions (blue) mapped onto a representative structure of a natural SH3 domain (from PI3K) bound to a proline-rich peptide ligand (pdb: 3I5R). (**B**) Results of statistical coupling analysis (SCA) on $n = 2593$ foldtuned SH3 sequences. Left: Second-order compressed coupling matrix, blocked into a single statistically interacting sector. Top right: First-order conservation scores. Bottom right: Visualization of sector positions (blue) mapped onto a representative structure of a natural SH3 domain (from PI3K) bound to a proline-rich peptide ligand (pdb: 3I5R)

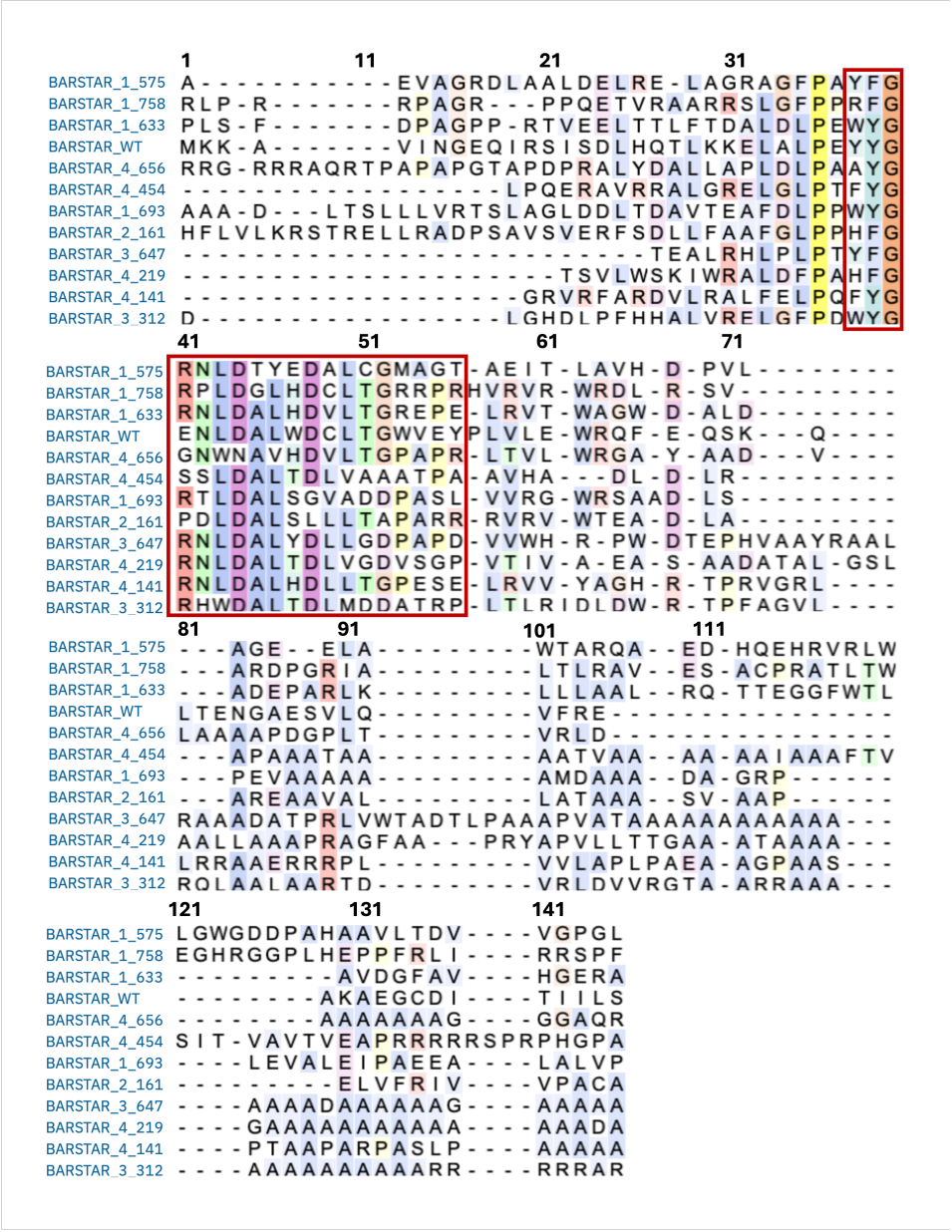

Figure S4.3: **Multiple sequence alignment (MSA) of toxicity-rescuing barstar variants.** Muliple sequence alignment (MSA) of the eleven toxicity-rescuing fold-tuned barstar variants and wild-type barstar from *B. aquaforiensis*. Columns corresponding to residue positions making physical contacts (positions 38-56; distance threshold < 4.0 Å) with barnase in a reference crystal structure (PDB: 1BRS) are boxed.

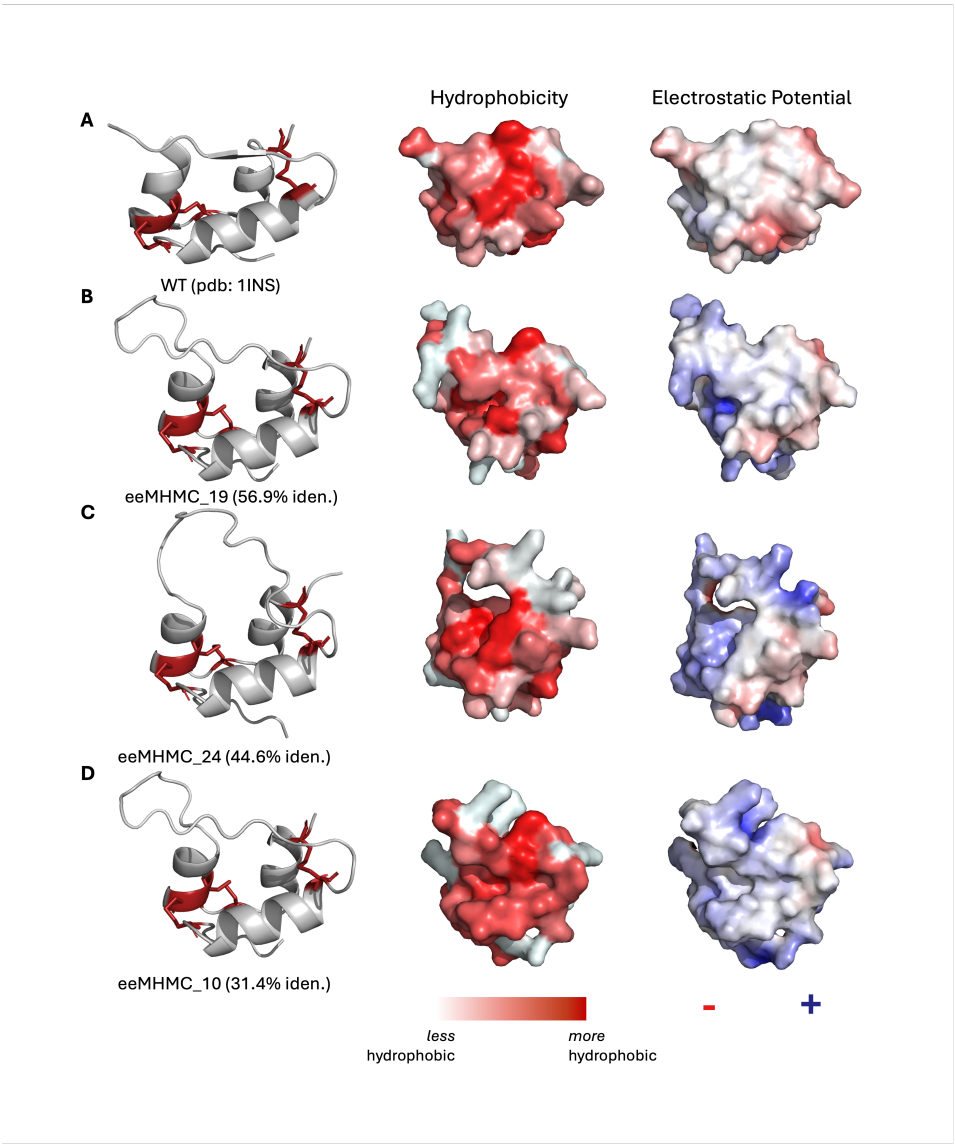

Figure S4.4: **eeMHMC-generated insulin variants respect structural and physicochemical plausibility of the canonical insulin peptide. left**: ESMFold predicted structures with disulfide bonds highlighted in red; **middle**: solvent-accessible surface visualization colored by hydrophobicity (Eisenberg scale); **right**: same as middle, colored by electrostatic potential with positive/negative charge in blue/red respectively. (**A**) wild-type human insulin (pdb: 1INS). (**B**) eeMHMC_19 (56.9% iden.). (**C**) eeMHMC_24 (44.6% iden.). (**D**) eeMHMC_10 (31.4% iden.)