Baker, E., Williams, C., Hudson, K., Bartlett, G., Heal, J., Porter Goff, K., ... Woolfson, D. (2017). Engineering protein stability with atomic precision in a monomeric miniprotein. *Nature Chemical Biology*, *13*(7), 764-770. https://doi.org/10.1038/nchembio.2380

Engineering protein stability with atomic precision

Emily G. Baker,[1]* Christopher Williams,[1,2,$] Kieran L. Hudson,[1,†,$] Gail J. Bartlett,[1] Jack W. Heal,[1] Kathryn L. Porter Goff,[1] Richard B. Sessions,[2,3] Matthew P. Crump[1,2] and Derek N. Woolfson[1,2,3]*

[1]School of Chemistry, University of Bristol, Cantock's Close, Bristol, BS8 1TS, UK
[2]BrisSynBio, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ
[3]School of Biochemistry, University of Bristol, Biomedical Sciences Building, University Walk, Bristol, BS8 1TD

[†]Current address: Department of Chemistry, UBC Faculty of Science, Vancouver Campus, 2036 Main Hall, Vancouver, BC Canada, V6T 1Z1

*Corresponding authors: EGB (emily.baker@bristol.ac.uk) and DNW (D.N.Woolfson@bristol.ac.uk)
[$]Contributed equally to this work.

One sentence summary:
Design and mutagenesis of a monomeric miniprotein provides insight into weak non-covalent interactions that help define and maintain folded proteins and protein-ligand interactions.

Abstract:
Miniproteins simplify the protein-folding problem, allowing the dissection of forces that stabilize protein structures. Here we describe PPα-Tyr, a designed peptide comprising an α helix buttressed by a polyproline-II helix. PPα-Tyr is water soluble, monomeric, and unfolds cooperatively with a midpoint unfolding temperature (T_M) of 39 ˚C. NMR structures of PPα-Tyr reveal proline residues docked between tyrosine side chains as designed. The stability of PPα is sensitive to the aromatic residue: replacing tyrosine by phenylalanine, *i.e.* changing three solvent-exposed hydroxyl groups to protons, reduces the T_M to 20 ˚C. We attribute this to the loss of CH–π interactions between the aromatic and proline rings, which we probe by substituting the aromatic residues with non-proteinogenic side chains. In analyses of natural protein structures we find a preference for proline-tyrosine interactions over other proline-containing pairs, and abundant CH–π interactions in biologically important complexes between proline-rich ligands and SH3 and similar domains.

The accumulation and cooperation of weak non-covalent interactions (NCIs) are critical for the stabilization of the folded, functional states of proteins.[1] In addition to hydrogen bonds, van der Waals' interactions and salt bridges, other NCIs are increasingly recognized as important contributors to protein stability, *e.g.,* CH–π, cation–π and $n \rightarrow \pi^*$ interactions.[2-5] Cooperativity, interplay, and even competition between many such weak interactions further complicate computational analysis and experimental dissection of NCIs.[6,7] Indeed, our current understanding of such forces and how they work together is incomplete and largely qualitative.

One route to improving our understanding of NCIs in proteins is to engineer or design smaller protein-like structures; *i.e.*, so-called miniproteins, which are polypeptide chains shorter than 40 – 50 amino acids with stable tertiary structures.[8-11] However, the requirement for optimized NCIs is even greater in these structures, where, despite the lower entropic cost of folding, the potential for NCIs is reduced because of their small size. Consequently, few miniproteins have been structurally characterized to high resolution, and of those that have many oligomerize,[12] are stabilized by disulfide bonds,[13-15] or depend on metal binding.[16]

For example, cysteine knots have two disulfide bonds that form a ring through which a third disulfide bond is threaded. This imparts exceptional stability even to enzymatic proteolysis.[14] The folding of zinc-finger peptides depends entirely on the binding of zinc, which is usually coordinated by sequence and spatially conserved cysteine and histidine residues. Remarkably, this leaves the majority of the remaining sequence positions free for mutations to many other amino acids without disrupting the overall tertiary structure. Although calcium-binding EF hands comprise two short α-helices separated by the metal-binding loop these usually dimerize for additional stability.[17]

The folded structures of the villin headpiece,[18] the tryptophan zipper,[19] the Trp-cage,[10] and most recently of TrpPlexus[11] are notable exceptions to the above, as the stabilities of these miniproteins are not contingent on the presence of covalent crosslinks or ligand binding. As the first example, the 20-residue Trp-cage peptide is particularly noteworthy. It has a short α helix that presents a single tryptophan (Trp) residue, which is penned in by three proline (Pro) residues from an abutting irregular piece of structure. The Trp-cage has a midpoint of thermal unfolding (T_M) of 42 ˚C, although the transition is broad and the peptide is fully folded only below 10 ˚C.[10] This stability has been improved by rational design.[20]

α Helices are standard building blocks in many natural proteins and the majority of successful protein designs described to date, including miniproteins. Although examples of persistent free-standing α helices are found in nature and have been designed,[21] α helices are usually stabilized through tertiary and quaternary interactions. Commonly, the α helices of natural and designed water-soluble proteins have *hpphppp* or similar sequence repeats of hydrophobic (*h*) and polar (*p*) residues. These patterns closely match the 3.6-residues-per-turn periodicity of the α helix, leading to amphipathic helices with distinct hydrophobic and polar faces. Stabilization is conferred through packing of the hydrophobic faces to those of other amphipathic secondary structures, leaving the polar faces exposed to aqueous media. The α-helical coiled coils are a well-understood example of this in which amphipathic helices oligomerize in prescribed ways that largely depend on the precise identities of the *h*-type residues.[12] More specifically, these residues project out from neighbouring helices and combine to give knobs-into-holes (KIH) packing, which define and stabilize the helical assemblies.[22,23] Short runs of these intimate KIH interactions can lead to a variety of very stable and specific quaternary structures formed by peptides just 20 – 40 residues in length.[24]

Here we describe the design and characterization of a series of short peptides, PPα, which adopt a stable monomeric fold that combines an amphipathic α helix and a stretch of polyproline-II helix. This compact tertiary structure is stabilized by tight inter-digitation of proline (Pro) residues from the latter and aromatic side chains displayed by the α helix. This packing is reminiscent of KIH interactions.[22,23] Our experimental studies of PPα and bioinformatics analyses of proline-aromatic side-chain contacts in protein structures more generally unveil a key role for CH–π interactions[3,25,26] in these Pro-Tyr-based packing arrangements. We argue that these contribute to the global stability of PPα, as well as for the affinities of proline-based protein-protein interactions more widely.

RESULTS

Miniprotein design and characterization

The design of PPα-Tyr borrowed from two natural proteins: a surface adhesin and antigen (AgI/II) from *Streptococcus mutans*, and the family of pancreatic polypeptide hormones (Supplementary Fig. 1).[27,28] In both structures, a polyproline-II helix and an α helix combine to form an unusual tertiary structure in which Pro residues from the former dock into holes formed by regularly spaced aromatic residues of the latter, Figures 1A-C. In AgI/II the helices are long, resulting in an overall fibrous structure, and the smaller peptide hormones form dimers.[29] To reduce this complexity—*i.e.*, to reduce the size of the system, and to eliminate dimer formation—we combined short segments of the two helical types *in silico*, choosing segment lengths to match best the different helical repeats, and sequences based on fragments from AgI/II, with tyrosine (Tyr) at the aromatic sites, Figure 1C. Engineering loops in protein design is notoriously difficult.[30,31] Therefore, we connected the two elements of secondary structure with the loop from the bovine pancreatic polypeptide hormone sequence. A full model for the resulting PPα-Tyr sequence, Table 1, with the topology polyproline-II helix—loop—α-helix, was constructed, energy minimized and found to be stable over 100 ns of molecular-dynamics (MD) simulations in water, Figure 1D.

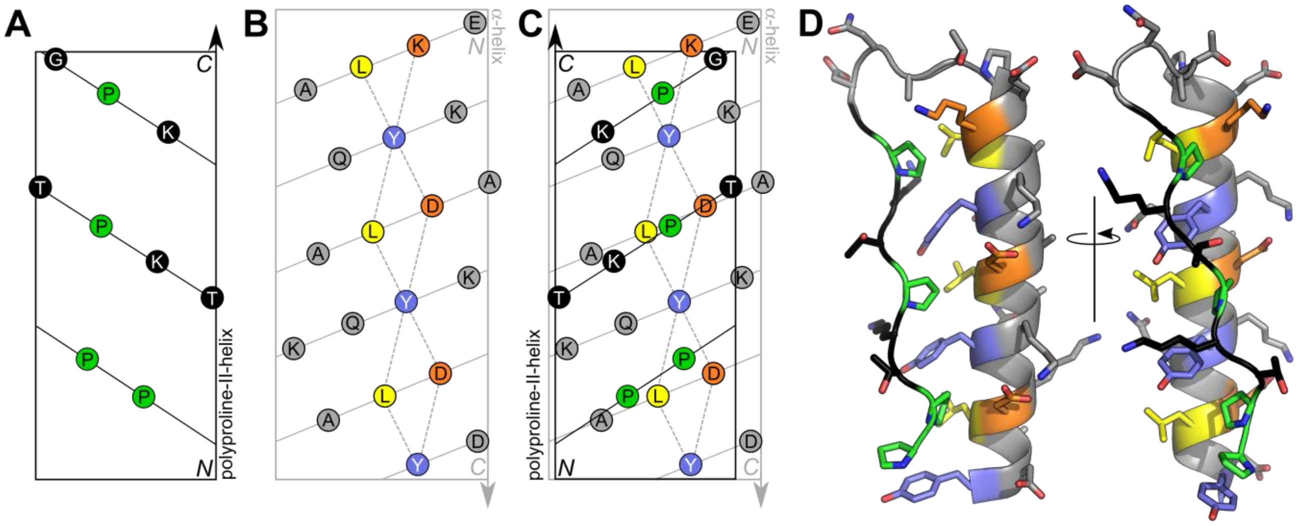

Figure 1: Design of PPα combining polyproline-II and α helices. A-C: 2D helical nets—*i.e.*, projections of Cα atoms onto the surfaces of cylinders of appropriate radii—for a canonical polyproline-II helix (**A**), an α helix (**B**), and with these two overlaid showing 'knobs-into-holes' packing of the Pro and Tyr side chains (**C**). The paths of the backbones are shown as solid lines, while dashed lines outline the 'holes' presented by the α helix. **Color key:** Tyr, slate; Leu, yellow; Lys and Asp, orange; and Pro, green. **D:** *In silico* model for the designed PPα-Tyr sequence, Table 1, after 100 ns of molecular-dynamics simulation in water.

Peptide	Sequence	AUC	T_M
	efgabcd efgabcd efgabcd	(x monomer)	(°C)
PPα-Tyr	Ac-PPTKPTKP GDNAT PEKLAKY QADLAKY QKDLADY-NH_2	0.9	39
PPα-Phe	Ac-PPTKPTKP GDNAT PEKLAKF QADLAKF QKDLADF-NH_2	0.9	20
PPα-Trp	Ac-PPTKPTKP GDNAT PEKLAKW QADLAKW QKDLADW-NH_2	0.9	36
PPα-His	Ac-PPTKPTKP GDNAT PEKLAKH QADLAKH QKDLADH-NH_2	ND	< 0
PPα-ϕNH_2	Ac-PPTKPTKP GDNAT PEKLAKϕ QADLAKϕ QKDLADϕ-NH_2	1.0	19
PPα-ϕOCH_3	Ac-PPTKPTKP GDNAT PEKLAKϕ QADLAKϕ QKDLADϕ-NH_2	1.0	38
PPα-ϕCH_3	Ac-PPTKPTKP GDNAT PEKLAKϕ QADLAKϕ QKDLADϕ-NH_2	0.8	31
PPα-ϕF	Ac-PPTKPTKP GDNAT PEKLAKϕ QADLAKϕ QKDLADϕ-NH_2	1.0	26
PPα-ϕCN	Ac-PPTKPTKP GDNAT PEKLAKϕ QADLAKϕ QKDLADϕ-NH_2	0.9	22
PPα-ϕNO_2	Ac-PPTKPTKP GDNAT PEKLAKϕ QADLAKϕ QKDLADϕ-NH_2	1.1	17

Table 1: Peptides designed and characterized in this study. Key: AUC, molecular weight relative to monomer mass from analytical ultracentrifugation; T_M, midpoint of thermal unfolding transition measured by CD spectroscopy; ϕ, non-proteinogenic amino acid based on L-phenylalanine with the *para* substituents given in the peptide name. Variants with ϕ = 4-trifluoromethyl-, 4-iodo-, 4-bromo- and 4-chloro-phenylalanine were also made, but these aggregated in solution.

A 34-residue synthetic peptide for PPα-Tyr (Supplementary Fig. 2) was soluble in aqueous buffer at pH 7.4. As judged by circular dichroism (CD) spectroscopy and consistent with the design, PPα-Tyr was folded with approximately 50% α-helical structure at 5 ˚C, Figure 2A and Supplementary Figure 4. Temperature dependence of the far-UV CD signal at 222 nm, which reports directly on the secondary structure present, revealed a reversible unfolding transition with a T_M of 39 ˚C, Figure 2B and Supplementary Figure 4. Furthermore, monitoring this transition by near-UV CD spectroscopy, which reports on the tertiary structure, gave an unfolding and refolding curve that were coincident with the far-UV CD traces (Supplementary Fig. 4). These data indicate fully cooperative unfolding and refolding behavior. Analytical ultracentrifugation (AUC) showed that PPα-Tyr was monomeric, Figure 2C and Supplementary Figure 5.

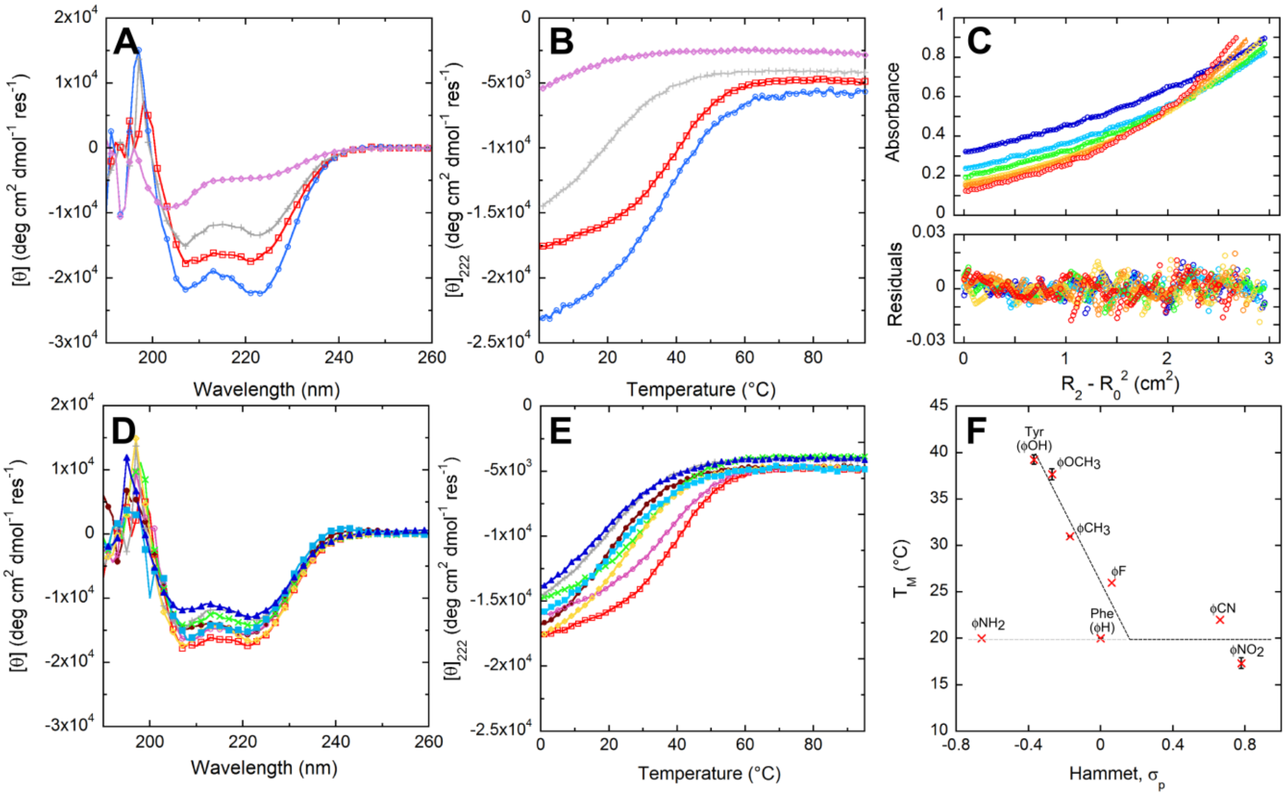

Figure 2: Folding and stability of PPα-Tyr and PPα variants. A: CD spectra recorded at 5 °C and **B:** thermal unfolding curves measured through the CD signal at 222 nm for PPα-Trp (blue circles), PPα-Tyr (red squares), PPα-Phe (gray crosses) and PPα-His (lilac diamonds). **C:** AUC data for PPα-Tyr (circles), and fits to a monomeric single ideal species (lines, upper panel), with residuals (lower panel) at rotor speeds of 40 krpm (blue), 44 krpm (light blue), 48 krpm (green), 52 krpm (yellow), 56 krpm (orange) and 60 krpm (red). **D:** CD spectra recorded at 5 °C for *p*-substituted phenylalanine-containing peptides; and **E:** thermal unfolding curves for the same peptides. **Color key for D&E:** (listed in order of σ_p values for the *p*-substituent) PPα-NH_2 (burgundy filled circles), PPα-Tyr (red squares), PPα-ϕOCH_3 (pink circles), PPα-ϕCH_3 (green saltires), PPα-Phe (gray crosses), PPα-ϕF (yellow diamonds), PPα-ϕCN (light blue filled squares) and PPα-ϕNO_2 (blue filled triangles). **F:** Plot of T_M values against the Hammett σ_p parameter for the corresponding aromatic substituent. Errors bars represent one standard deviation from the mean of at least three data sets. Dashed lines are included simply to guide the eye. In parts A, B, D&E, representative spectra from at least three replicate experiments are shown.

High-resolution and high-sensitivity nuclear magnetic resonance (NMR) spectroscopy was used to determine the solution structure of PPα-Tyr. This employed standard homo-nuclear experiments and natural-abundance ^{15}N- and ^{13}C-edited HSQC spectra. 87% of the ^{1}H resonances were assigned, Supplementary Table 1, with the side chains of solvent-exposed lysine residues mostly accounting for the missing assignments. Consistent with the design, the PPα-Tyr structure comprised a polyproline-II helix and loop (residues 1—13) and an α helix (residues 14—33), Figures 3A&B. The core of the structure was highly defined with numerous strong NOEs between the aromatic side chains and surrounding residues. Unsurprisingly, the conformations of some of the solvent-exposed side chains were less well defined and could not be fully assigned, which resulted in some variation across the ensemble: the backbone RMSD was 0.514 Å ± 0.121 Å, and the all-atom RMSD 0.825 ± 0.122 Å, Supplementary Figure 6. A representative structure from this ensemble matched the *in silico* model with RMSDs of 0.7 Å and 1.3 Å measured over the backbone and all atoms, respectively. Moreover, at the helix-helix interface KIH-type packing between the tyrosine and proline residues was evident, and these side chains were in close contact, Figure 3A&B.

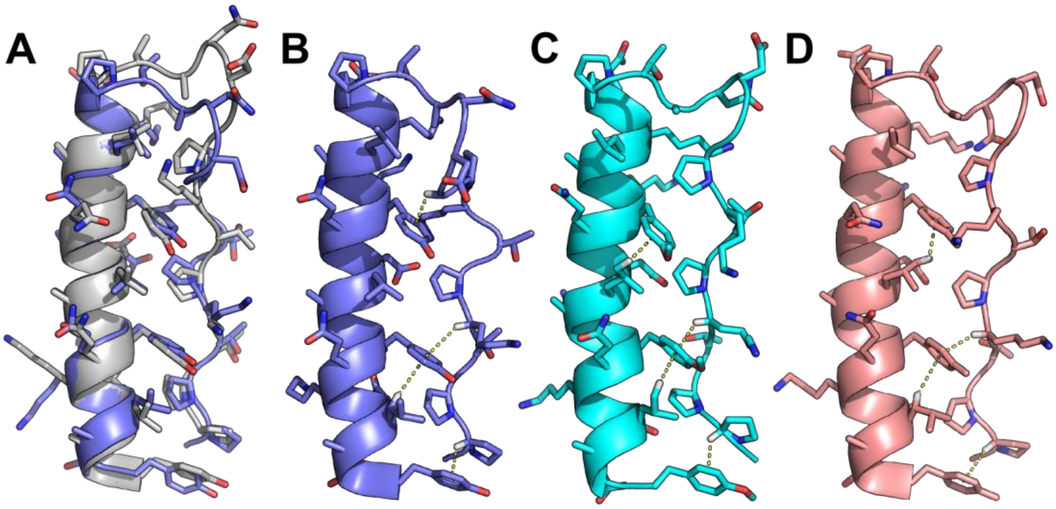

Figure 3: NMR structures for the *p*-substituted phenylalanine variants of PPα. A: NMR structure closest to the geometric mean of the ensemble (model 20) for PPα-Tyr (slate) overlaid with the *in silico* model after 100 ns of MD (gray). **B-D:** Representative NMR structures from the ensembles showing the CH–π interactions found for PPα-Tyr, model 14 (**B**), PPα-ϕOCH_3, model 8 (**C**), and PPα-ϕCH_3, model 5 (**D**). The average numbers of CH–π interactions per ensemble structure were 2.25, 2.7 and 2.5, respectively, with 1.2, 0.65 and 0.55 per structure involving Pro. Although the remaining PPα peptides were folded by NMR they gave poor-quality spectra and structure calculations were not possible, which corroborated their reduced thermal stability. PDB codes: PPα-Tyr, 5LO2; PPα-ϕOCH3, 5LO3; and PPα-ϕCH3, 5LO4.

Intimate Pro-aromatic contacts

Because of the close contacts between the tyrosine and proximal aliphatic side chains, we searched for potential CH–π interactions in PPα-Tyr. To do this, we used an operational definition for these interactions and parameters adapted from previous studies (Supplementary Fig. 8).[3,32] We found 24 CH–π interactions between these residues across the ensemble of 20 structures, and detected additional CH–π interactions involving 15 lysine, 4 leucine and 2 glycine residues as the CH donors, Supplementary Table 2.

On this basis, we posited that the stability of PPα might be promoted through improved CH–π interactions with the Tyr residues substituted for tryptophan (Trp), which has a more electron-rich aromatic system, and reduced when replaced by histidine (His) or phenylalanine (Phe), which have less electron-rich rings.[2,32] Consistent with this, but nevertheless surprisingly, PPα-His was largely unfolded as judged by CD spectroscopy, Figs 2A&B. We took the characterization of this peptide no further. The other variants, PPα-Trp and PPα-Phe, Table 1, were soluble, cooperatively folded by CD spectroscopy (Figs. 2A&B and Supplementary Fig. 4), and monomeric in AUC, (Supplementary Fig. 5). PPα-Phe was destabilized to a significant degree with the T_M reduced to 20 °C compared with 39 °C for PPα-Tyr. This is remarkable given the small chemical changes involved, *i.e.* three solvent-exposed hydroxyl groups in the protein were replaced by protons. Contrary to initial expectations, the stability of PPα-Trp (T_M = 36 °C) was comparable to that of PPα-Tyr. Thus, whilst electron-poor His and Phe do destabilize PPα, both electron-rich aromatics (Tyr and Trp) stabilize the structure to similar extents.

To understand this better, we analyzed interactions between Pro and aromatic side chains, and, for comparison, all aliphatic—aromatic side-chain contacts in the RCSB Protein Data Bank (PDB), Figure 4 and Supplementary Figure 8. We used only non-redundant (<40% sequence identity) X-ray protein crystal structures of 1 Å resolution or better, and those that had all CH protons experimentally determined. A side-chain contact map for the propensity of interactions between Val, Ile, Leu, Pro, Phe, Tyr and Trp, revealed several trends, Figure 4A, some of which have been noted by others:[3,33] (1) like-with-like contacts were favored, *i.e.* aliphatic—aliphatic and aromatic—aromatic; (2) in general, aliphatic—aromatic contacts were neutral, *i.e.* they occurred at rates expected by chance; and (3) Phe was unusual in that it made more contacts than expected with all of the other hydrophobic residues except Pro. However, we found Pro broke these patterns: despite having an aliphatic cyclic side chain, it contacted the other aliphatic residues and Phe less often than expected; and, in contrast, Pro interacted with Tyr and Trp significantly more often than expected by chance, Fig. 4A and Supplementary Fig. 7.

Closer examination of the Pro-aromatic pairings from the PDB showed that approximately one quarter had potential CH–π interactions, Supplementary Table 3. Moreover, Pro-Trp and Pro-Tyr pairs participated in these predicted CH–π interactions much more than Pro-Phe, Figure 4B. As a result, these pairs were highly directional compared with the more-isotropic distributions of aliphatic residues around aromatics, compare Figures 4C&D. This directionality likely arises from a combination of electrostatic and electronic interactions between the electron-rich aromatic groups, and the slightly acidic protons of the pyrrolidine ring of Pro, which are both consistent with electrostatic surface potentials, Figures 4E-G and Supplementary Figure 9.

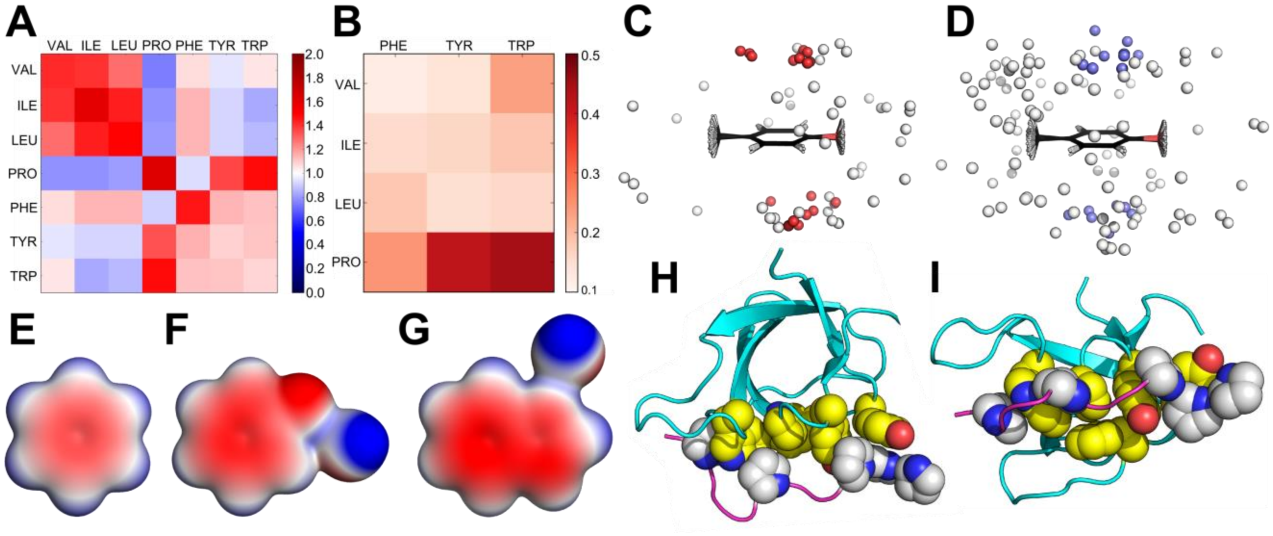

Figure 4: Pairwise side-chain and CH–π interactions in the PDB. A: Heat map for the propensity (observed/expected ratio) of amino-acid pairs with one or more sub-3 Å atom-atom contacts. **B:** Heat map of the proportion of aliphatic-aromatic close contacts that participate in CH–π interactions normalized for propensity of the pairs to be in close contact. **C:** Overlay of Pro-Tyr side-chain contacts within 3 Å; gray spheres represent the centers of mass of Pro side chains, with those that tested positive for CH–π interactions colored red. **D:** Similar to **C** but for Val-Tyr contacts, and CH–π positive interactions colored slate. **E-G:** Electrostatic surface potentials (ESPs) of Phe (**E**), Tyr (**F**), and Trp (**G**) side chains. Tyr and Trp are shown as hydrogen-bond donors to a water molecule to best represent their solvated state. Scale: ≤ -130 kJ mol^{-1} (electropositive, blue) through ≥ 130 kJ mol^{-1} (electronegative, red). **H&I:** Orthogonal views of human adapter protein Tuba SH3 domain (PDB: 4CC7), which has 7 CH–π interactions between its binding domain and Pro-rich peptide of N-WASP. **Color key**: atoms of proline side chains of the ligands, gray and blue; atoms of the interacting side chains from the SH3 domain, yellow, red and blue.

Non-proteinogenic substitutions in PPα

To probe CH–π interactions in the PPα system, ten further variants were synthesized with *para*-substituted phenylalanine residues at the aromatic sites covering electron-rich *p*-methoxyphenylalanine through electron-poor *p*-nitrophenylalanine, Table 1 and Supplementary Figure 9. Six of the peptides were soluble, folded, monomeric, and gave full or near-complete thermal unfolding curves, Figures 2D&E and Supplementary Figures 4&5. NMR structures for the most-stable variants, *p*-methoxyphenylalanine and *p*-methylphenylalanine, again revealed intimate contacts and CH–π interactions between the pyrrolidine ring of Pro and the faces of modified aromatic rings (Figs. 3C&D). The numbers of contacts made consistent with CH–π interactions across the 20 conformers of these two ensembles were 54 and 50, respectively.

To probe the contribution of these potential CH–π interactions to PPα stability, the stabilities of the *para*-substituted phenylalanine variants were plotted against the corresponding Hammett constant σ_p, Figure 2F. Formally, the Hammett equation relates the equilibrium constant for the dissociation of substituted benzoic acids to two parameters: the substituent or Hammett constant, σ; and the reaction constant, ρ. The Hammett constant provides a measure of how much the substituent stabilizes the negative charge of the conjugate base. Traditionally, it is interpreted in terms of through-bond inductive and mesomeric effects that alter the electrostatics of the ring. Whilst we recognize that this has potential caveats,[34] here we use σ_p as a proxy for the electron density in the aromatic ring,[35] Supplementary Fig. 9, to compare the thermal unfolding reactions of the PPα variants. On this premise, we plotted the T_M's of each mutant against σ_p for the appropriate substituted aromatic residue Figures 2F, and we also plotted various thermodynamic parameters obtained from fitting of the full unfolding curves against σ_p (Supplementary Fig. 10).

The data for PPα-Tyr, PPα-ϕOMe, PPα-ϕCH_3, PPα-ϕF and PPα-Phe were close to linear with a negative slope. This is strong evidence for electrostatic and electronic contributions to aromatic-Pro interactions over and above the hydrophobic effect and van der Waals' interactions.[36] Specifically, it provides evidence for CH–π interactions, which would redistribute electron density from the ring into the CH bond, and so be favored by the electron-donating groups in this series. The Hammett plot leveled off for the PPα-ϕCN and PPα-ϕNO_2 variants, consistent with arguments that cyano- and nitro-functionalized benzenes have weaker interaction energies with XH groups[37] and consequently, weaker CH–π interactions in PPα.

n.b. The stability of PPα-ϕNH_2 was lower than expected based on the σ_p value of aniline. We have no clear explanation for this. We measured the pK_a of the *p*-amino group in the peptide using a pH titration and following the UV spectrum, but we found it was unperturbed from that of the free amino acid. Thus, we assume that the lone pair of electrons of the substituent is fully available to the π-system and that the σ_p value is appropriate. Because of the reduced stability of this peptide compared to PPα-Tyr, a full assignment of the NMR signals was not possible nor was a structure determination.

The thermal denaturation profiles of Figure 2E were fitted by van't Hoff analyses to determine ΔH_{unf}, ΔS_{unf} and ΔG_{unf} at 5 ˚C where all of the peptides were close to fully folded, Supplementary Figure 10 and Supplementary Table 4. Interpreting ΔS_{unf} and ΔH_{unf} values for protein folding is complicated. Therefore, we focused on the free energies of unfolding, ΔG_{unf}, which differed between the mutants, and, like the T_M values (Fig. 2F), these varied linearly with σ_p, Supplementary Figure 10. The ΔG_{unf} values were spread over 3.6 kJ mol^{-1} ≈ 0.9 kcal mol^{-1}. With 2 – 3 CH–π interactions per structure from the NMR data, it is interesting that this energy is close to literature estimates for CH–π interactions of ≈ 1.5 – 2.8 kcal mol^{-1}.[38] Though small, energy differences on this scale shift equilibrium or binding constants by nearly an order of magnitude. Thus, the presence of even a small number of these NCIs will influence the energetics of biomolecular folding and association considerably.

Pro-aromatic interactions in ligand binding

With the potential contributions to free energies of binding in mind, we examined Pro-aromatic contacts known to be important in natural biological processes. Specifically, interactions between SH3, WW, EVH1 and profilin domains and their target proline-rich ligands were inspected, Figures 4H&I.[39,40] Amongst other functions, these protein-peptide interactions control pathways in cell growth, transcription, cytoskeletal remodeling and other regulatory functions across all kingdoms of life. Within the 596 X-ray crystal and NMR structures containing such domains in the PDB, 135 chains had non-covalently bound polypeptide ligands with ≥ 3 contiguous residues in polyproline-II-helix conformations, Supplementary Table 5. When culled at 80% protein-sequence identity, and taking only X-ray crystal structures of ≤ 2.1 Å resolution along with NMR structures, this yielded 38 complexes. On average, the polyproline-II stretches of the ligands in the assessed structures were 4 – 5 residues long. Within this set, there were 121 CH–π inter-chain interactions, at an average of 3.18 CH–π interactions per complex. 55% of Pro, which accounted for 149 of 407 ligand residues, participated in CH–π interactions. This is significantly more than the 16% of Pro that form CH–π interactions across the entire PDB, Supplementary Tables 3&5. In other words, Pro-aromatic and CH–π interactions in the SH3 and similar domains are denser and more frequent than those generally found in proteins. Tyr was the most frequent CH–π partner for Pro in these protein-ligand interactions, followed by the two rings of Trp, and then Phe, Supplementary Figure 8.

DISCUSSION

In conclusion, we report the fragment-based design and complete structural characterization of a new miniprotein, PPα, with a stable, monomeric polyproline-II helix—loop—α-helix fold. In the design, the lengths of the two helices were chosen to best match the different repeats of the two types of helix. This was done to promote intimate knobs-into-holes packing of Pro and Tyr side chains from the polyproline-II and the α helix, respectively. Our biophysical data and high-resolution solution-phase NMR structures validate this approach. Moreover, they reveal that, over and above the anticipated hydrophobic effect and van der Waals' forces from the packing arrangement, PPα is stabilized by CH–π interactions between the Pro and Tyr side chains. This is supported by stability studies in a series of *para*-substituted phenylalanine mutants of PPα, which confirm an electrostatic/electronic component to the Pro-aromatic interactions: peptides with electron-rich aromatic π-systems are more thermally stable and have more favorable free-energies of folding than those with electron-withdrawing substituents. Of the proteinogenic aromatic amino acids, the electron-rich Tyr and Trp give more stable PPα folds and appear to make better CH–π interactions than the Phe and His mutants.

Analyses of the RCSB Protein Data Bank add considerably to these conclusions: Pro-Tyr and Pro-Trp interactions are observed much more frequently than expected by chance, and also more frequently than any other aliphatic-aromatic side-chain pairings. By contrast, Pro-Phe contacts are underrepresented. Furthermore, Pro-Tyr and Pro-Trp make many more CH–π interactions than any of the other side-chain interactions. More specifically, protein-ligand interactions involving proline-rich ligands, such as those found in SH3 domains, indicate that Pro-Tyr contacts are particularly favored and lead to unusually high densities of CH–π interactions in these complexes. This is noteworthy because the literature on protein-peptide interactions of this type focuses on the stabilizing influence of only the hydrophobic effect. Therefore, we propose that CH–π interactions also contribute to the observed affinities of these short linear-peptide ligands.

Our observations raise the question: why does Pro interact preferably with Tyr rather than the larger π system of Trp in these cases? We suggest that the single aromatic ring of Tyr allows sufficient Pro-aromatic contacts, whereas the larger Trp makes packing more difficult and may even lead to lower solubility of the unbound states of the ligand. The last point could be important for both systems examined herein where the aromatic residues are partly exposed, as in the adhesins and pancreatic polypeptides,[27,28] or exposed part of the time, as with the free SH3 and other domains.[41]

Our findings indicate that CH–π interactions, which are traditionally considered as weak NCIs, can have considerable impact on protein folding and stability. Moreover, these interactions could be particularly important in the design and optimization of miniproteins, protein mimics, protein-ligand interactions and possibly catalysts.[42] Therefore, and as we have shown, unpicking the contributions of such NCIs to protein stability, folding and association using the subtleties of non-proteinogenic side chains will be critical in developing our understanding of and for manipulating these fundamental forces.[43,44] As one of the smallest, monomeric globular protein folds described to date, PPα provides a particularly attractive model system for advancing such studies. Finally, we encourage others to consider weak NCIs, such as CH–π interactions, in the design and development of small molecules that mimic or disrupt currently undruggable natural protein-protein interactions.[13,45]

Acknowledgments:
EGB and DNW are supported by a BBSRC/ERASynBio grant (BB/M005615/1); KLH, GJB, JWH and DNW are supported by the ERC (340764); DNW is a Royal Society Wolfson Research Merit

Award holder (WM140008); and KLPG is supported by the EPSRC-funded Bristol Chemical Synthesis Centre for Doctoral Training (EP/G036764/1). We thank BrisSynBio for access to the BBSRC/EPSRC-funded 700 MHz NMR spectrometer (BB/L01386X/1), S. Whittaker and the Henry Wellcome Building NMR Facility at the University of Birmingham for access to the Wellcome Trust–funded 900 MHz spectrometer (099185/Z/12/Z), and R. Alder and members of the Woolfson group for helpful discussions.

Author contributions:

EGB and DNW designed the research. EGB made the synthetic peptides and performed the CD spectroscopy and AUC experiments. CW and MPC collected the NMR data. CW, KLPG and EGB analyzed the NMR data, and CW solved the NMR structures. KLH, GJB, DNW conducted the bioinformatics. EGB and RBS carried out the MD studies. JWH and EGB performed the van't Hoff analyses. EGB and DNW wrote the paper. All authors reviewed and contributed to the manuscript.

References:

1 Pace, N.C., Scholtz, J.M. & Grimsley, G.R. Forces Stabilizing Proteins. *FEBS Lett.* **588**, 2177-2184 (2014).

2 Dougherty, D.A. Cation-π Interactions in Chemistry and Biology: A New View of Benzene, Phe, Tyr, and Trp. *Science* **271**, 163-168 (1996).

3 Brandl, M., Weiss, M.S., Jabs, A., Sühnel, J. & Hilgenfeld, R. C-H···π-Interactions in Proteins. *J. Mol. Biol.* **307**, 357-377 (2001).

4 Bartlett, G.J., Choudhary, A., Raines, R.T. & Woolfson, D.N. n→π* Interactions in Proteins. *Nat. Chem. Biol.* **6**, 615-620 (2010).

5 Yesselman, J.D., Horowitz, S., Brooks III, C.L. & Trievel, R.C. Frequent Side Chain Methyl Carbon-Oxygen Hydrogen Bonding in Proteins Revealed by Computational and Stereochemical Analysis of Neutron Structures. *Proteins: Struct. Funct. Bioinf.* **83**, 403-410 (2015).

6 Bhattacharya, A., Tejero, R. & Montelione, G.T. Evaluating Protein Structures Determined by Structural Genomics Consortia. *Proteins: Struct. Funct. Bioinf.* **66**, 778-795 (2007).

7 Bartlett, G.J., Newberry, R.W., VanVeller, B., Raines, R.T. & Woolfson, D.N. Interplay of Hydrogen Bonds and n→π* Interactions in Proteins. *Journal of the American Chemical Society* **135**, 18682-18688 (2013).

8 Zondlo, N.J. & Schepartz, A. Highly Specific DNA Recognition by a Designed Miniature Protein. *J. Am. Chem. Soc.* **121**, 6938-6939 (1999).

9 Gellman, S.H. & Woolfson, D.N. Mini-Proteins Trp the Light Fantastic. *Nat. Struct. Biol.* **9**, 408-410 (2002).

10 Neidigh, J.W., Fesinmeyer, R.M. & Andersen, N.H. Designing a 20-Residue Protein. *Nat. Struct. Mol. Biol.* **9**, 425-430 (2002).

11 Craven, T.W., Cho, M.-K., Traaseth, N.J., Bonneau, R. & Kirshenbaum, K. A Miniature Protein Stabilized by a Cation−π Interaction Network. *J. Am. Chem. Soc.* **138**, 1543-1550 (2016).

12 Woolfson, D.N. The Design of Coiled-Coil Structures and Assemblies. *Adv. Protein Chem.* **70**, 79-112 (2005).

13 Golemi-Kotra, D. *et al.* High Affinity, Paralog-Specific Recognition of the Mena EVH1 Domain by a Miniature Protein. *J. Am. Chem. Soc.* **126**, 4-5 (2004).

14 Daly, N.L. & Craik, D.J. Bioactive Cystine Knot Proteins. *Curr. Opin. Chem. Biol.* **15**, 362-368 (2011).

15 Bhardwaj, G. *et al.* Accurate de novo Design of Hyperstable Constrained Peptides. *Nature* **538**, 329-335 (2016).

16 Pabo, C.O., Peisach, E. & Grant, R.A. Design and Selection of Novel Cys(2)His(2) Zinc Finger Proteins. *Annu. Rev. Biochem.* **70**, 313-340 (2001).

17 Gifford, J.L., Walsh, M.P. & Vogel, H.J. Structures and Metal-Ion-Binding Properties of Ca^{2+}-Binding Helix-Loop-Helix EF Hand Motifs. *Biochem. J.* **405**, 199-221 (2007).

18 McKnight, C.J., Matsudaira, P.T. & Kim, P.S. NMR Structure of the 35-Residue Villin Headpiece Subdomain. *Nat. Struct. Biol.* **4**, 180-184 (1997).

19 Cochran, A.G., Skelton, N.J. & Starovasnik, M.A. Tryptophan Zippers: Stable, Monomeric β-Hairpins. *Proc. Natl. Acad. Sci. USA* **98**, 5578-5583 (2001).

20 Barua, B. *et al.* The Trp-Cage: Optimizing the Stability of a Globular Miniprotein. *Prot. Eng. Des. Sel.* **21**, 171-185 (2008).

21 Baker, E.G. *et al.* Local and Macroscopic Electrostatic Interactions in Single α-Helices. *Nat. Chem. Biol.* **11**, 221-228 (2015).

22 Crick, F.H.C. The Packing of α-Helices: Simple Coiled-Coils. *Acta Crystallogr.* **6**, 689-697 (1953).

23 Walshaw, J. & Woolfson, D.N. SOCKET: A Program for Identifying and Analysing Coiled-Coil Motifs Within Protein Structures. *J. Mol. Biol.* **307**, 1427-1450 (2001).

24 Woolfson, D.N. *et al.* De novo Protein Design: How Do We Expand into the Universe of Possible Protein Structures? *Curr. Opin. Struct. Biol.* **33**, 16-26 (2015).

25 Plevin, M.J., Bryce, D.L. & Boisbouvier, J. Direct Detection of CH/π Interactions in Proteins. *Nat. Chem.* **2**, 466-471 (2010).

26 Nishio, M., Umezawa, Y., Fantini, J., Weiss, M.S. & Chakrabarti, P. CH-π Hydrogen Bonds in Biological Macromolecules. *Phys. Chem. Chem. Phys.* **16**, 12648-12683 (2014).

27 Blundell, T.L., Pitts, J.E., Tickle, I.J., Wood, S.P. & Wu, C.-W. X-Ray-Analysis (1.4-Å Resolution) of Avian Pancreatic-Polypeptide: Small Globular Protein Hormone. *Proc. Natl. Acad. Sci. USA* **78**, 4175-4179 (1981).

28 Larson, M.R. *et al.* Elongated Fibrillar Structure of a Streptococcal Adhesin Assembled by the High-Affinity Association of α- and PPII-Helices. *Proc. Natl. Acad. Sci. USA* **107**, 5983-5988 (2010).

29 Noelken, M.E., Chang, P.J. & Kimmel, J.R. Conformation and Association of Pancreatic-Polypeptide from 3 Species. *Biochemistry* **19**, 1838-1843 (1980).

30 Fiser, A., Do, R.K.G. & Šali, A. Modeling of Loops in Protein Structures. *Prot. Sci.* **9**, 1753-1773 (2000).

31 Hu, X.Z., Wang, H.C., Ke, H.M. & Kuhlman, B. High-Resolution Design of a Protein Loop. *Proc. Natl. Acad. Sci. USA* **104**, 17668-17673 (2007).

32 Hudson, K.L. *et al.* Carbohydrate-Aromatic Interactions in Proteins. *J. Am. Chem. Soc.* **137**, 15152-15160 (2015).

33 Saha, R.P., Bhattacharyya, R. & Chakrabarti, P. Interaction Geometry Involving Planar Groups in Protein-Protein Interfaces. *Proteins: Struct. Funct. Bioinf.* **67**, 84-97 (2007).

34 Wheeler, S.E. & Houk, K.N. Through-Space Effects of Substituents Dominate Molecular Electrostatic Potentials of Substituted Arenes. *J. Chem. Theory Comput.* **5**, 2301-2312 (2009).

35 J. Carver, F., A. Hunter, C. & M. Seward, E. Structure-Activity Relationship for Quantifying Aromatic Interactions. *Chem. Commun.*, 775-776 (1998).

36 Zondlo, N.J. Aromatic-Proline Interactions: Electronically Tunable CH/π Interactions. *Acc. Chem. Res.* **46**, 1039-1049 (2013).

37 Bloom, J.W.G., Raju, R.K. & Wheeler, S.E. Physical Nature of Substituent Effects in XH/π Interactions. *J. Chem. Theory Comput.* **8**, 3167-3174 (2012).

38 Tsuzuki, S., Honda, K., Uchimaru, T., Mikami, M. & Tanabe, K. The Magnitude of the CH/π Interaction between Benzene and Some Model Hydrocarbons. *J. Am. Chem. Soc.* **122**, 3746-3753 (2000).

39 Kay, B.K., Williamson, M.P. & Sudol, P. The Importance of Being Proline: The Interaction of Proline-Rich Motifs in Signaling Proteins with Their Cognate Domains. *FASEB J.* **14**, 231-241 (2000).

40 Kay, B.K. SH3 Domains Come of Age. *FEBS Lett.* **586**, 2606-2608 (2012).

41 Ball, L.J., Kuhne, R., Schneider-Mergener, J. & Oschkinat, H. Recognition of Proline-Rich Motifs by Protein-Protein-Interaction Domains. *Angew. Chem. Int. Ed.* **44**, 2852-2869 (2005).

42 Parsons, Z.D., Bland, J.M., Mullins, E.A. & Eichman, B.F. A Catalytic Role for C–H/π Interactions in Base Excision Repair by *Bacillus cereus* DNA Glycosylase AlkD. *J. Am. Chem. Soc.* **138**, 11485-11488 (2016).

43 Zhang, Y.T., Malamakal, R.M. & Chenoweth, D.M. Aza-Glycine Induces Collagen Hyperstability. *J. Am. Chem. Soc.* **137**, 12422-12425 (2015).

44 Arnold, U. & Raines, R.T. Replacing a Single Atom Accelerates the Folding of a Protein and Increases its Thermostability. *Org. Biomol. Chem.* **14**, 6780-6785 (2016).

45 Cobos, E.S. *et al.* A Miniprotein Scaffold Used to Assemble the Polyproline II Binding Epitope Recognized by SH3 Domains. *J. Mol. Biol.* **342**, 355-365 (2004).