

## **A glutamine-based single $\alpha$ -helix scaffold to target globular proteins**

A. Escobedo<sup>\*,1</sup>, J. Piccirillo<sup>1</sup>, J. Aranda<sup>1</sup>, T. Diercks<sup>2</sup>, B. Topal<sup>1</sup>, M. Biesaga<sup>1</sup>, L. Staby<sup>3</sup>, B. B. Kragelund<sup>3</sup>, J. García<sup>1</sup>, O. Millet<sup>2</sup>, M. Orozco<sup>1,4</sup>, M. Coles<sup>5</sup>, R. Crehuet<sup>6</sup> and X. Salvatella<sup>\*,1,7</sup>

<sup>1</sup>Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldori Reixac 10, 08028 Barcelona, Spain.

<sup>2</sup>CIC bioGUNE, Bizkaia, Science and Technology Park bld 801A, 48160 Derio, Bizkaia, Spain.

<sup>3</sup>REPIN and Structural Biology and NMR Laboratory, The Linderstrøm-Lang Centre for Protein Science, Department of Biology, Ole Maaloes Vej 5, University of Copenhagen, DK-2200 Copenhagen N, Denmark.

<sup>4</sup>Department of Biochemistry and Biomedicine, University of Barcelona, Avinguda Diagonal 645, 08028 Barcelona, Spain.

<sup>5</sup>Department of Protein Evolution, Max Planck Institute for Developmental Biology, Max-Planck-Ring 5, 72076 Tübingen, Germany

<sup>6</sup>CSIC-Institute for Advanced Chemistry of Catalonia (IQAC), Jordi Girona 18-26, 08034 Barcelona, Spain.

<sup>7</sup>ICREA, Passeig Lluís Companys 23, 08010, Barcelona, Spain

\*to whom correspondence should be addressed: [albert.escobedo@crg.eu](mailto:albert.escobedo@crg.eu),  
[xavier.salvatella@irbbarcelona.org](mailto:xavier.salvatella@irbbarcelona.org)

Current address:

A. Escobedo, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain.

J. Piccirillo, Department of Macromolecular Structures, Centro Nacional de Biotecnología (CNB-CSIC), Madrid, Spain.

## **Abstract**

**The binding of intrinsically disordered proteins to globular ones often requires the folding of motifs into  $\alpha$ -helices. These interactions offer opportunities for therapeutic intervention but their modulation with small molecules is challenging because they bury large surfaces. Linear peptides that display the residues that are key for binding can be targeted to globular proteins when they form stable helices, which in most cases requires their chemical modification. Here we present rules to design peptides that fold into single  $\alpha$ -helices by instead concatenating glutamine side chain to main chain hydrogen bonds recently discovered in polyglutamine helices. The resulting peptides are uncharged, contain only natural amino acids, and their sequences can be optimized to interact with specific targets. Our results provide design rules to obtain single  $\alpha$ -helices for a wide range of applications in protein engineering and drug design.**

## Introduction

Proteins are central to biology as they carry out a wide range of essential functions, from gene regulation to enzymatic catalysis, where their ability to specifically interact with other biomolecules is key. In pharmacology, inhibiting their interactions using drug-like small molecules is a common approach to modulate biological functions relevant to disease. In cases where the binding partner is another protein, the binding interfaces are usually flat and extended<sup>1</sup>, making it challenging to inhibit the interaction with small molecules<sup>2</sup>, and it is generally preferable to target them with antibodies. Yet, despite recent remarkable progress in intracellular antibody delivery<sup>3</sup>, their clinical applications have been limited to targeting extracellular proteins, highlighting the need to develop new molecular tools to inhibit protein-protein interactions.

Peptides have some of the advantages of small molecules, such as their ease of synthesis and rather high bioavailability, and some of the advantages of antibodies, such as their relatively large size. Peptides, therefore, have at least in principle great potential as modulators of protein-protein interactions for pharmacological applications<sup>4,5</sup>. Protein-protein interactions where one partner adopts a helical conformation upon binding are particularly common and especially amenable to inhibition by peptides: an excised linear peptide comprising a suitable sequence can, in principle, inhibit the interaction if it can bind to its partner with high affinity<sup>6</sup>. Linear peptides have a low propensity to fold into stable  $\alpha$ -helices, however, and the entropic cost of folding decreases both their affinity for their targets and their stability against proteolytic degradation, highlighting the need to develop new tools to stabilize their helical conformation.

Introducing non-natural amino acids with reactive side chains and connecting them via covalent bonds, as in peptide stapling<sup>7,8</sup>, is a powerful method to constrain linear peptides into conformations akin to  $\alpha$ -helices. Stapling is based on the pairwise use of synthetic  $\alpha$ -methyl,  $\alpha$ -alkenyl amino acids at relative positions  $i,i+3$ ,  $i,i+4$  or  $i,i+7$  where their side chains can react by ring closing metathesis when the peptide folds into an  $\alpha$ -helix, thus greatly stabilizing this conformation; in some cases, two such connections were concatenated to yield particularly long and stable helices<sup>9</sup>. While such stapled  $\alpha$ -helical peptides have shown broad applicability to inhibit protein-protein interactions<sup>10</sup>, they have drawbacks that limit their range of applicability, such as their costly synthesis, limited solubility, and high rigidity.

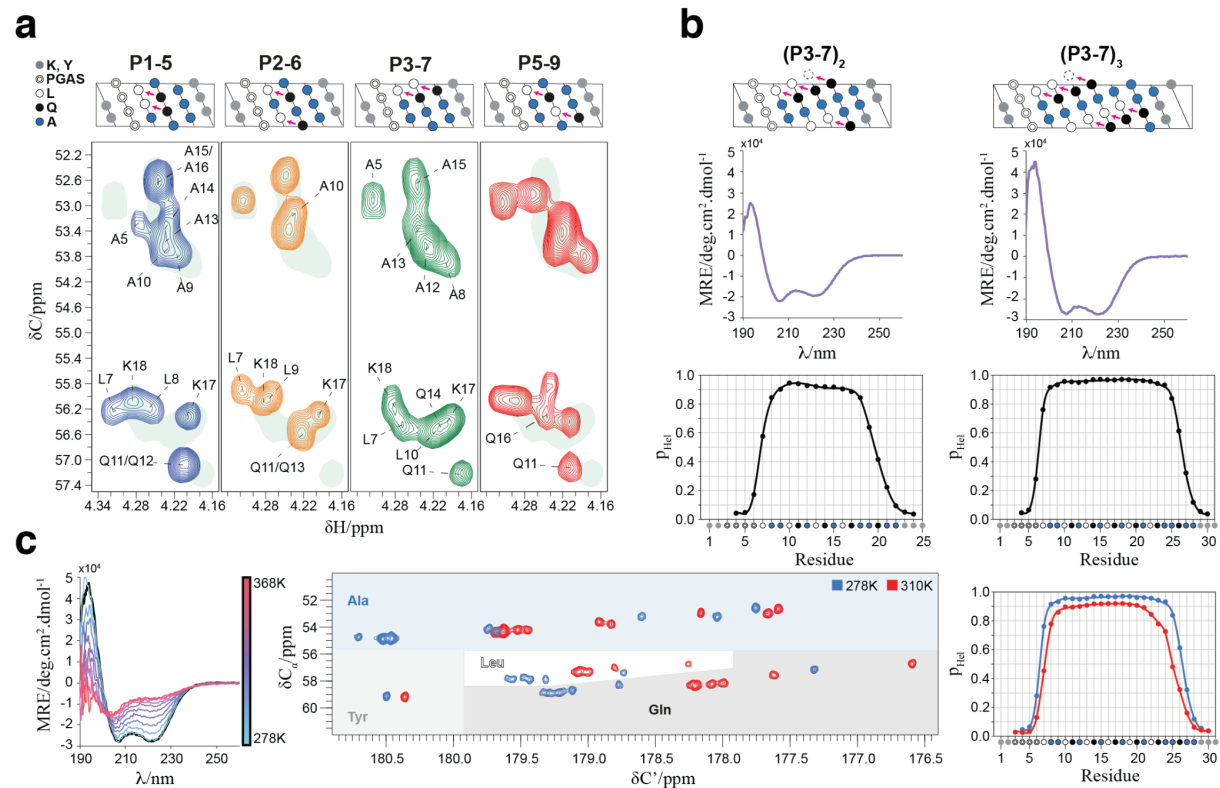
We recently reported that Gln side chain to main chain hydrogen bonds stabilize the

polyglutamine (polyQ) helix, a helical secondary structure formed by the polyQ tract of the androgen receptor (AR)<sup>11,12</sup>. In this structure both the main and side chain amide groups of Gln residues at position  $i$  donate hydrogens to the main chain CO group of the residues at position  $i-4$ . The strength of this bifurcate interaction depends on the residue at position  $i-4$ , with Leu performing particularly well. This interaction has also been observed in the polyQ tract of the protein huntingtin, indicating that it is not specific to AR and can therefore be used in design<sup>13</sup>. As the relative donor and acceptor positions of such interactions are equivalent to those of the covalently linked residues in stapled peptides, we reasoned that they would stabilize the  $\alpha$ -helical conformation of linear peptides to enhance their interaction with globular proteins.

To address this hypothesis we design a series of linear, uncharged, and highly soluble peptides and confirm their cooperative folding into single  $\alpha$ -helices under physiological conditions. To evaluate the versatility and range of applicability of the design rules we analyze their tolerance towards changes of the acceptor residue, with bulky hydrophobic residues performing best, and introduce a pH-dependent conformational switch; in addition we explore how such bifurcated hydrogen bonds can be combined with electrostatic interactions between side chains to further stabilize helical structure. Most importantly, we show that the sequences of such peptides can be tailored to interact with specific globular proteins. In summary, the simple design rules that we propose can be used to engineer a new class of linear, cooperatively folded helical peptides for use as templates in applications in pharmacology, materials science, synthetic biology and, more generally, in bioengineering.

## Results

### Design of Gln-based single $\alpha$ -helices



**Fig. 1 | Design of single  $\alpha$ -helices stabilized by Gln side chain to main chain hydrogen bonds.** **a** Top: representation, as helical projections, of the peptides used to investigate cooperativity, where pink arrows indicate putative side chain to main chain interactions. Bottom: selected region of the 2D  $^1\text{H}$ ,  $^{13}\text{C}$  HSQC spectra of the peptides at 278 K, where the shaded area corresponds to that of peptide P3-7. **b** Top: representation, as helical projections, of peptides  $(\text{P3-7})_2$  and  $(\text{P3-7})_3$ , as in panel a. Middle: CD spectra at 278 K. Bottom: residue-specific helical propensities as obtained from the  $\text{C}'$ ,  $\text{C}_\alpha$ ,  $\text{N}^{\text{H}}$ , and  $\text{H}^{\text{N}}$  NMR chemical shifts using CheSPI<sup>14,15</sup>. **c** Left: CD spectra of peptides  $(\text{P3-7})_2$  and  $(\text{P3-7})_3$  obtained at temperatures between 278 K (blue) and 368 K (red) in steps of 10 K; the spectrum obtained at 278 K after refolding is shown in black. Center: superposition of a selected region of the  $^{13}\text{C}$  detected 2D CACO NMR spectra of peptide  $(\text{P3-7})_3$  recorded at 278 K (blue) and 310 K (red), with an indication of amino acid specific regions as shades. Right: residue-specific helical propensities as obtained from the NMR chemical shifts at 278 and 310 K as in **b**.

Structure prediction algorithms are powerful tools for protein design. Machine learning based algorithms predict the structures of globular proteins<sup>16</sup> but their applicability to proteins devoid of tertiary interactions is still relatively unexplored<sup>17</sup>. This is illustrated by the results reported for intrinsically disordered (ID) domains possessing polyglutamine (polyQ) tracts in the AlphaFold protein structure database. These tracts are predicted to fold into  $\alpha$ -helices<sup>17,18</sup> whereas experiments indicate that their helical propensity strongly depends on sequence context<sup>11,13,19–22</sup>. The helical propensity of peptides and ID proteins can instead be predicted

by algorithms, such as Agadir, that compute the energetics of folding by considering residue helical propensities, capping effects and interactions between side chains<sup>23</sup>. Agadir does not account for the Gln side chain to main chain interactions stabilizing polyQ helices and therefore underestimates the helicity of the polyQ tract in AR: the peptide L<sub>4</sub>Q<sub>16</sub>, excised from AR, has 38% helical propensity according to NMR experiments while Agadir predicts only 3%<sup>11</sup>. To address this we introduced an additional energy term accounting for this interaction ( $E_{i,i+4}^L$ ) and by minimizing the prediction error ( $RMSD_{Hel}$ ) obtained  $E_{i,i+4}^L = -0.6$  kcal mol<sup>-1</sup> for L<sub>4</sub>Q<sub>16</sub>, in the range expected for one hydrogen bond in water<sup>24</sup>. Peptides with a short polyQ tract did not require this correction term (Fig. S1a), however, and  $E_{i,i+4}^L$  depended on tract length: it increased from -0.4 kcal mol<sup>-1</sup> for L<sub>4</sub>Q<sub>8</sub> to -0.7 kcal mol<sup>-1</sup> for L<sub>4</sub>Q<sub>20</sub>. The effective strength of the Gln<sub>*i*+4</sub>→Leu<sub>*i*</sub> interactions in AR peptides depends thus on the number of equivalent interactions following them in the sequence, indicating cooperativity.

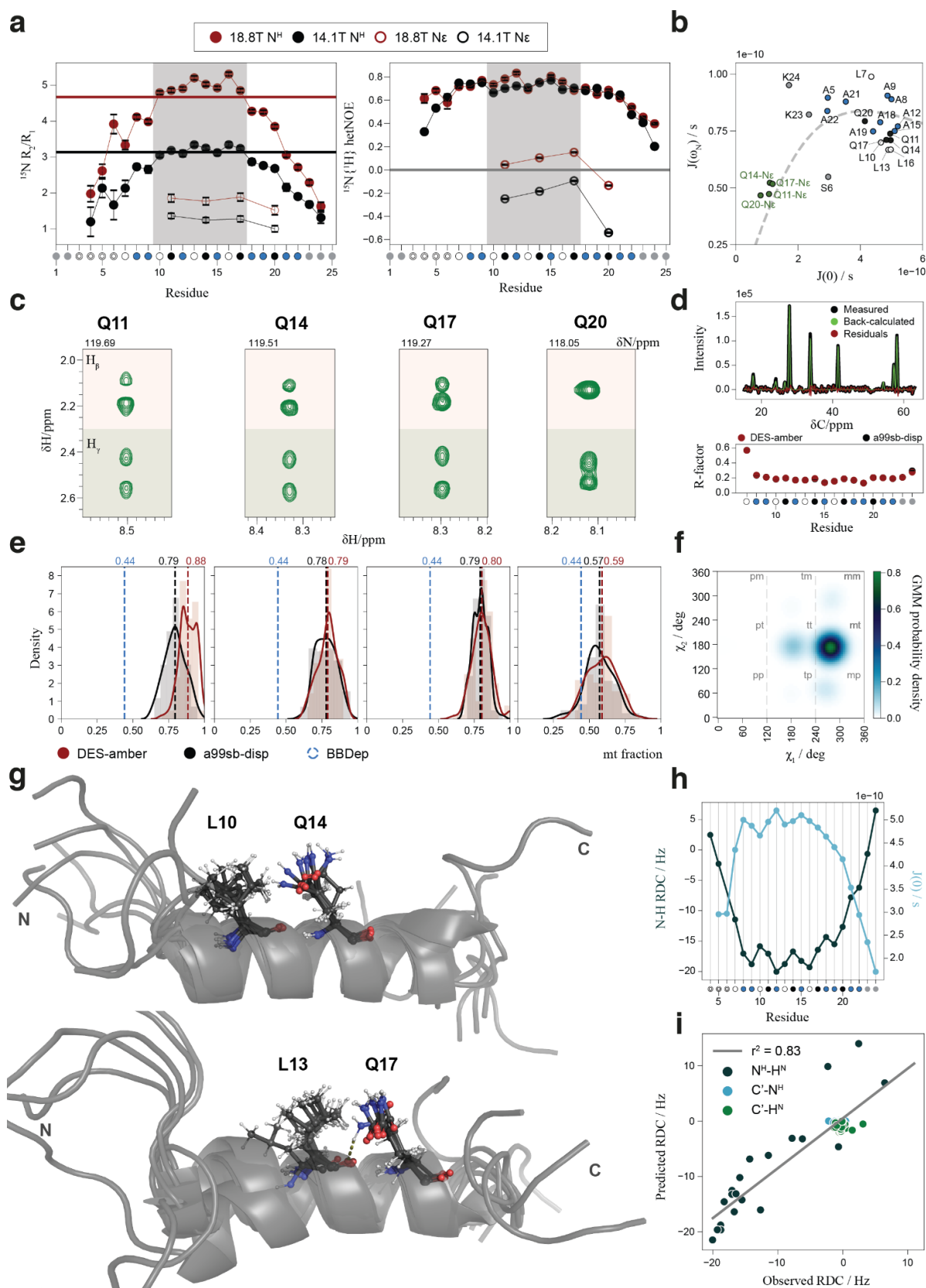
To analyze the origin of this behavior and exploit it for peptide design, we studied four peptides of identical amino acid composition but with two potential Gln<sub>*i*+4</sub>→Leu<sub>*i*</sub> interactions (pink arrows in Fig. 1a) at different relative positions. The first such interaction is common to all peptides and the second one is shifted 1 (peptide P1-5), 2 (P2-6), 3 (P3-7) or 5 (P5-9) positions towards the C-terminus. After confirming that they were monomeric under our experimental conditions by size exclusion chromatography coupled to multiple angle light scattering (SEC-MALS) (Fig. S1d), we used solution state nuclear magnetic resonance (NMR) spectroscopy to probe their conformation by exploiting the dependence of <sup>13</sup>C<sub>α</sub> and <sup>1</sup>H<sub>α</sub> NMR chemical shifts on secondary structure, where larger <sup>13</sup>C<sub>α</sub> and lower <sup>1</sup>H<sub>α</sub> shifts indicate higher helicity<sup>25</sup>. The NMR spectra showed P3-7 is the most helical peptide (Fig. 1a), indicating that the strength of two Gln<sub>*i*+4</sub>→Leu<sub>*i*</sub> interactions is maximal when the donor of the first (Gln11 in P3-7) and the acceptor of the second (Leu10) interaction share a peptide bond such that the two interactions are concatenated; these results were confirmed by circular dichroism (CD) spectroscopy (Fig. S1c).

We then used CD spectroscopy to study the secondary structure of peptides containing two and three pairs of concatenated Gln<sub>*i*+4</sub>→Leu<sub>*i*</sub> interactions, (P3-7)<sub>2</sub> and (P3-7)<sub>3</sub>, and obtained that they are highly helical and monomeric ( $\Theta_{222\text{ nm}}/\Theta_{208\text{ nm}} < 1$ ) (Fig. 1b). An analysis of secondary structure based on main chain NMR chemical shifts indicated that the residue-specific helical propensity ( $p_{Hel}$ ) is larger than 0.9 across 9 contiguous residues for (P3-7)<sub>2</sub> and larger than 0.95 across 16 residues for (P3-7)<sub>3</sub>. As it is not common for monomeric peptides to cooperatively fold into α-helices in the absence of tertiary interactions, we verified their monomeric state by SEC-MALS and native mass spectrometry (MS) (Figs. S1d and S1e). In addition we removed the N-terminal PGAS motif, that can

facilitate helix nucleation<sup>11</sup>, from these peptides (Fig. 1a) and observed that the *uncapped* counterparts,  $u(P3-7)_2$  and  $u(P3-7)_3$ , also have high helical propensity (Fig. S1f).

Finally we investigated the thermal stability of the helices by CD spectroscopy at temperatures up to 368 K (Fig. 1c). The CD spectra at 278 K were equivalent to those obtained upon cooling after thermal unfolding, indicating that the unfolded state of the peptide is soluble under our experimental conditions. We also characterized the structural properties of  $(P3-7)_3$  by NMR at physiological temperature, 310 K. A comparison of the <sup>13</sup>C-detected 2D CACO spectra of  $(P3-7)_3$  at 278 and 310 K revealed only a small decrease in helical propensity (Fig. 1c, center and right) indicating that the peptide remains essentially fully folded at 310 K ( $p_{\text{Hel}} \approx 0.90$  over 12 contiguous residues). In conclusion, concatenating  $\text{Gln}_{i+4} \rightarrow \text{Leu}_i$  interactions allows obtaining stable  $\alpha$ -helices that remain folded under physiological conditions, are soluble upon thermal unfolding and only contain natural amino acids.

## Structure of a Gln-based single $\alpha$ -helix



**Fig. 2 | Structure of a Gln-based single  $\alpha$ -helix.** **a**  $^{15}\text{N}$  NMR relaxation data for peptide  $(\text{P3-7})_2$  main chain ( $\text{N}^{\text{H}}$ ) and Gln side chain ( $\text{N}_{\epsilon}$ ) amide groups, measured at 14.1 T (black) and 18.8 T (dark red) at 278 K in 10%  $\text{D}_2\text{O}$  (main chain) and 50%  $\text{D}_2\text{O}$  (side chain), where residues are labeled as in Fig. 1. Left:  $^{15}\text{N}$   $R_2/R_1$  ratios. Solid horizontal lines represent the theoretical  $R_2/R_1$  ratio of a rigid body in isotropic motion with  $\tau_c = 4.5$  ns, whose field dependency agrees well with the experimental data. Right:  $^{15}\text{N}\{^1\text{H}\}$  heteronuclear NOE, where the shade corresponds to the region with the highest degree of structuration. **b**. Spectral densities,  $J(\omega)$ , derived by reduced spectral density mapping of all  $^{15}\text{N}$  relaxation data: the dashed line corresponds to the behavior expected for a polypeptide. **c** Strips from a 3D  $^{15}\text{N}$ -edited TOCSY-HSQC spectrum showing the  $\text{H}_{\beta}$  and  $\text{H}_{\gamma}$  resonances of all four Gln residues in peptide  $(\text{P3-7})_2$ . **d** Residue-specific CoMAND fitting of NOE signal intensities from the 3D CNH-NOESY (i.e.  $[\text{H}]\text{C}, \text{NH}$  HSQC-NOESY-HSQC) spectrum of peptide  $(\text{P3-7})_2$ . Top: comparison of measured and back-calculated (using CoMAND and the a99sb-disp trajectory frame pool) 1D  $^{13}\text{C}$  traces for the Gln14 main chain N. Bottom: Average R-factors obtained from 100 CoMAND iterations using either DES-amber or a99sb-disp trajectory frame pools. **e** Density plots of the fraction of the *mt* rotamer found in the 100 fitting iterations for all four Gln residues. The fraction of helical glutamines in the *mt* configuration in the BBDep dataset<sup>26</sup> is shown in blue and marked with a blue dashed line, whereas black and dark red values and the corresponding dashed lines indicate the mean *mt* content for the a99sb-disp and DES-amber distributions, respectively. **f** Probability density for Gln14  $\chi_1$  and  $\chi_2$  torsion angles derived by CoMAND using the Gaussian mixture model (GMM). **g** Structural ensemble of  $(\text{P3-7})_2$ , with main chain conformations selected by CoMAND from both trajectory pools and side chain conformations generated by GMM sampling. Top: one of the 20 global ensemble calculations aligned for the Leu 10 Gln14 pair. Bottom: same iteration aligned for the Leu13 Gln17 pair. **h** Residual dipolar couplings (RDCs) for the main chain  $^1\text{H}-^{15}\text{N}$  moieties (dark blue, left axis) and relaxation-derived main chain spectral density at zero frequency,  $J(0)$ , both showing the dipolar wave modulation characteristic of  $\alpha$ -helices. **i** Correlation between experimental and back-calculated RDCs, computed as the average of the ensemble produced by pooling the ensembles obtained in 20 independent iterations.

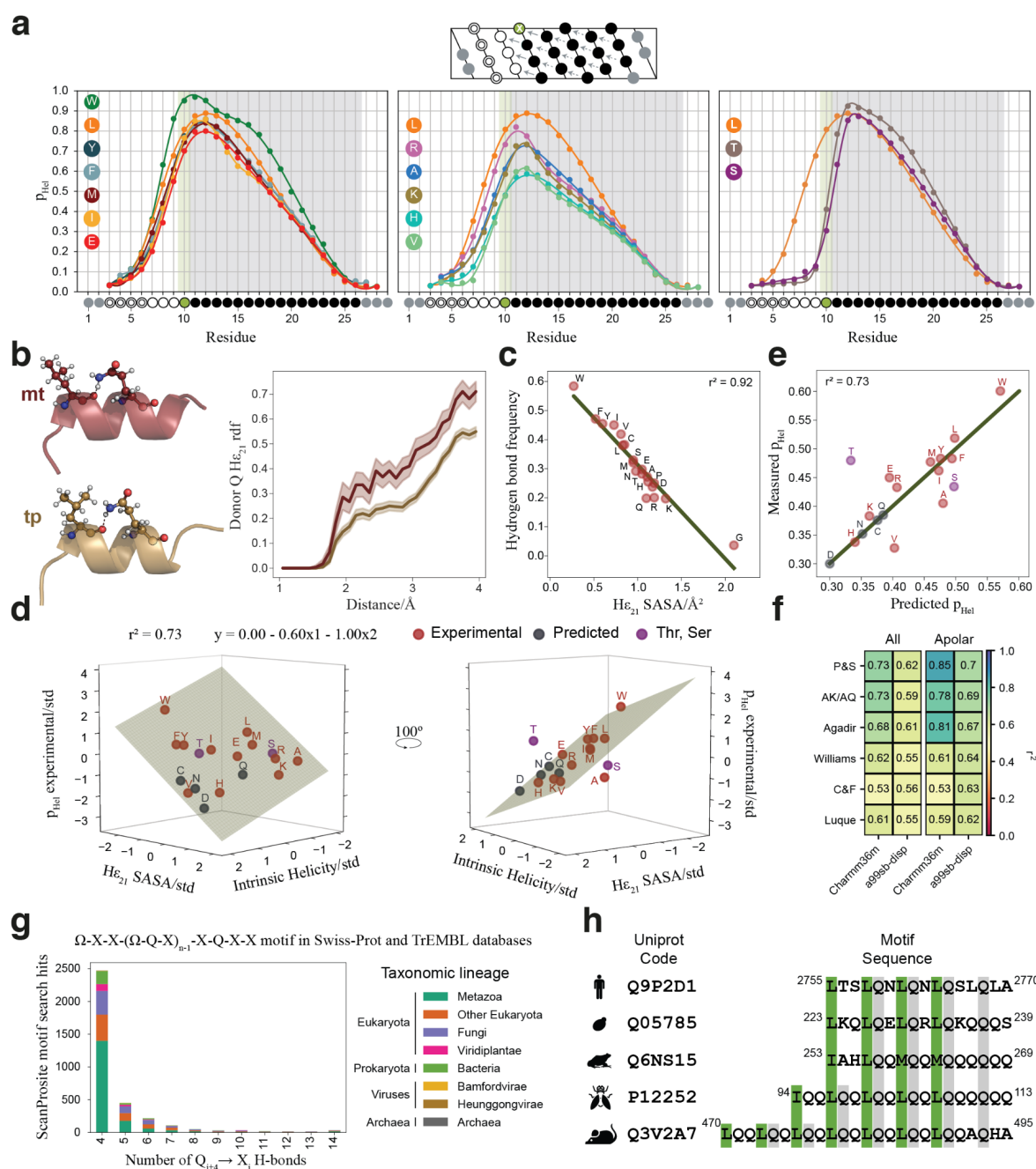
The high quality of the NMR spectra obtained for peptide  $(\text{P3-7})_2$  allowed using this technique to study its structure at atomic resolution and further characterize the interactions stabilizing its conformation (Fig. S2a). First we measured  $^{15}\text{N}$  relaxation at 278 K ( $R_1$ ,  $R_2$ , heteronuclear  $^{15}\text{N}\{^1\text{H}\}$  NOE) at two magnetic field strengths (14.1 T and 18.8 T) for the main chain amide (NH) groups of residues Gly4 to Lys24 and for the side chain amide ( $\text{N}_{\epsilon 2}\text{H}_{\epsilon 21}$ ) groups of all four Gln residues (Gln11, Gln14, Gln17, Gln20) (Fig. 2a, S2b). We found that the main chain  $^{15}\text{N}$   $R_2/R_1$  ratios increase from the termini towards the center of the peptide until reaching plateau values ( $5.05 \pm 0.25$  at 18.8 T,  $3.20 \pm 0.15$  at 14.1 T) between residues Leu10 and Gln17 (Fig. 2a). A similar trend is traced out by the  $^{15}\text{N}\{^1\text{H}\}$  NOE, that reaches upper plateau values between 0.65 and 0.83 over a larger central segment, from Leu7 to Ala18. The main chain amide  $^{15}\text{N}$  relaxation data thus localizes the region of highest structural rigidity on a  $10^{-1}$  to  $10^1$  ns timescale as the one defined by the pairs of Leu and Gln residues involved in concatenated  $\text{Gln}_{i+4} \rightarrow \text{Leu}_i$  interactions i.e. from Leu10 to Gln17. By contrast, the  $\text{Gln20} \rightarrow \text{Leu16}$  interaction, involving the last Gln residue, appears to be weaker.

The different properties of the first three Gln residues (11, 14 and 17) relative to the last Gln residue (20) are also evident in the side chain amide  $^{15}\text{N}_{\epsilon 2}$  relaxation data: while positive (at 18.8 T) or close to zero (at 14.1 T)  $^{15}\text{N}_{\epsilon 2}\{^1\text{H}_{\epsilon 21}\}$  NOE values suggest significantly less side chain mobility for the first three Gln residues, Gln20 shows clearly negative values in line with higher side chain mobility (Fig. 2a), also reflected in the relaxation-derived spectral density map (Fig. 2b). These results agree with the notion that the first three Gln residues form stronger  $\text{Gln}_{i+4} \rightarrow \text{Leu}_i$  interactions than Gln20. The high NMR signal dispersion in both the  $^{13}\text{C}$  and  $^{15}\text{N}$  dimensions of the spectra of peptide  $(\text{P3-7})_2$  also allowed studying the Gln side chains: the first three Gln residues show fully resolved  $\text{H}_\beta$  and  $\text{H}_\gamma$  signals in the  $^{15}\text{N}$ -edited HSQC-TOCSY spectrum, whereas the  $\text{H}_\beta$  signals overlap for Gln 20 (Fig. 2c). This indicates that the conformations of the side chains of the former are better defined than those of the latter, even more than in polyQ helices<sup>11,13</sup>, in agreement with the relaxation data.

To improve the description of the Gln side chains, we used CoMAND<sup>27</sup> to infer rotamer populations from a diagonal-free 3D CNH-NOESY spectrum reporting on distances between protons bound to  $^{15}\text{N}$  and  $^{13}\text{C}$ . Briefly, CoMAND selects subsets of conformers from a pool, here obtained by molecular dynamics, that reproduce the NOESY spectra (Fig. 2d). The distributions were enriched in the *mt* Gln rotamer ( $\chi_1 = -60^\circ$  and  $\chi_2 = 180^\circ$ ) that is required for the  $\text{Gln}_{i+4} \rightarrow \text{Leu}_i$  interaction: the Gln residues involved in strong interactions (11, 14 and 17) have a high population (0.80), whereas Gln 20 has a lower population (0.58). Both values are higher than that obtained for Gln residues in  $\alpha$ -helices of structures deposited in the PDB<sup>26</sup>, 0.44 (Fig. 2e); these results were robust to changes in the force field used to generate the pool<sup>28,29</sup>.

To generate a conformational ensemble, we used the residue-specific CoMAND ensembles to train a Gaussian mixture model (GMM) by inferring  $\chi_1$  and  $\chi_2$  probability densities for each residue (Figs. 2f and S2c), modified the side chain conformations accordingly and, through R-factor minimization, obtained the set of representative conformers shown in Figure 2g. To validate it we measured three sets of residual dipolar couplings (RDCs) under steric alignment. The main chain  $^1\text{D}_{\text{H,N}}$  values show a dipolar wave pattern typical of helices that matches well the period of the zero frequency spectral density,  $J(0)$  (Fig. 2h). Although the RDCs were not used as restraints they correlate well with the ensemble-averaged values ( $Q = 0.37$ , Fig. 2i), confirming that the ensemble is an accurate representation of peptide  $(\text{P3-7})_2$  and that the design rules that we put forward lead to single  $\alpha$ -helices.

# Gln side chain to main chain hydrogen bonds can be accepted by residues other than Leu



**Fig. 3 | Gln side chain to main chain hydrogen bonds can be accepted by different residues. a** Top:

Representation, as helical projection, of the  $L_3XQ_{16}$  variants studied in this work with the color code described in Fig. 1. Bottom: residue-specific helical propensity of the  $L_3XQ_{16}$  variants derived from the backbone  $N$ ,  $H^N$ ,  $C_\alpha$  and  $C'$  chemical shifts. The type of residue  $X$  (in position 10) is indicated by the colored circles. Left: helical profile of the 6 most helical single variants ( $X = W, Y, F, M, I, E$ ) compared to that of  $L_4Q_{16}$ . Center: helical profile of the least helical single variants ( $X = R, A, K, H, V$ ) compared to that of  $L_4Q_{16}$ . Right: Helical profile of the outlier variants ( $X = T, S$ ) compared to that of  $L_4Q_{16}$  (spectra in Fig. S3b). **b** Effect of the rotameric state of the acceptor on the interaction of atom  $H_{e21}$  with  $H_2O$ . Left: two frames of the Charmm36m 1  $\mu s$  MD trajectory obtained for  $L_4Q_8$  showing Gln4 involved in a bifurcated hydrogen bond with Leu4 in either the *mt* (top) or *tp* (bottom) rotamer of the

Leu residue. Right: radial distribution function for Gln4 H<sub>ε21</sub> in the frames where Leu4 populates the *mt* (red) or *tp* (gold) rotamer. **c** The frequency of the side chain to main chain hydrogen bond is strongly correlated with the solvent accessibility surface area (SASA) of H<sub>ε21</sub>, which depends on type of residue X. **d** Multiple regression correlating intrinsic helicity (x1) (Pace and Scholtz scale<sup>30</sup>) and SASA (x2) with the average helicity (y): the data were standardized to estimate the relative weight of each variable in defining the model. The experimental data points used for regression are shown in dark red, the outliers in purple, and the model-predicted data points in teal. **e** Measured *versus* predicted average L<sub>3</sub>XQ<sub>16</sub> helicity. Color code as in **d**. **f** Squared r correlation scores (r<sup>2</sup>) for the multiple regression shown in **e**, using different reported scales for intrinsic amino acid helicity (y axis) and H<sub>ε21</sub> SASA values derived from two sets of 1 μs MD trajectories independently generated with different force fields. The results are shown for all experimentally measured variants (left) and those with an apolar residue (L, I, V, F, Y, M, W, A) in position X (right). **g** Number of ScanProsite-identified protein sequences in UniprotKB (including the Swiss-Prot and TrEMBL databases) containing the (P3-7)<sub>n</sub> motif with an increasing number of Q<sub>i+4</sub> → Ω<sub>i</sub> pairs (Ω-X-X-(Ω-Q-X)<sub>n-1</sub>-X-Q-X-X motif search), where Ω represents any good side chain to main chain hydrogen bond acceptor W, L, Y, F, I and M, whereas X represents any amino acid. Counts are grouped by taxonomic lineage as obtained from UniprotKB. **h** Representative natural sequences that fulfill the design rules identified in proteins with UniprotKB annotation score = 5 and for which there is evidence at the protein level, belonging to a variety of organisms (from top to bottom: *Homo sapiens*, *Saccharomyces cerevisiae*, *Xenopus laevis*, *Drosophila melanogaster*, *Mus musculus*).

The stability of the Gln-based α-helices stems from concatenated Gln side chain to main chain hydrogen bonds accepted by Leu residues. This design decision was based on the fact that the polyQ tract found in AR, which is the most helical studied so far, is flanked by four Leu residues<sup>11</sup>. We sought to determine how other residues perform as acceptors to increase the versatility of the design rules and better understand the factors determining the strength of the interaction. For this we used a host-guest approach in which we determined the secondary structure of L<sub>3</sub>XQ<sub>16</sub> peptides (Table S1) by NMR. These peptides were obtained by mutating the fourth Leu residue of peptide L<sub>4</sub>Q<sub>16</sub>, excised from AR, to 13 representative amino acids (Fig. 3a).

We measured the residue-specific helical propensities of the peptides by NMR by combining standard <sup>1</sup>H<sup>N</sup>-detected triple resonance with <sup>13</sup>C-detected CACO and CON 2D NMR experiments: the high resolution in the CO dimension of the latter allowed the unambiguous assignment of all Gln residues, even in the variants with lowest signal dispersion (Fig. S3b). Except L<sub>3</sub>TQ<sub>16</sub> and L<sub>3</sub>SQ<sub>16</sub> all variants show a helicity profile approximately proportional to that of L<sub>4</sub>Q<sub>16</sub>, (Fig. 3a): L<sub>3</sub>TQ<sub>16</sub> and L<sub>3</sub>SQ<sub>16</sub> instead show a different profile, likely because in these cases the substitution shifts the site of helix nucleation by introducing an S/T N-capping motif<sup>31</sup>. In these, Ser/Thr accept two concomitant hydrogen bonds donated by main chain amides of residues C-terminal to them: one by the main chain O and another one by the side chain O of the hydroxyl group.

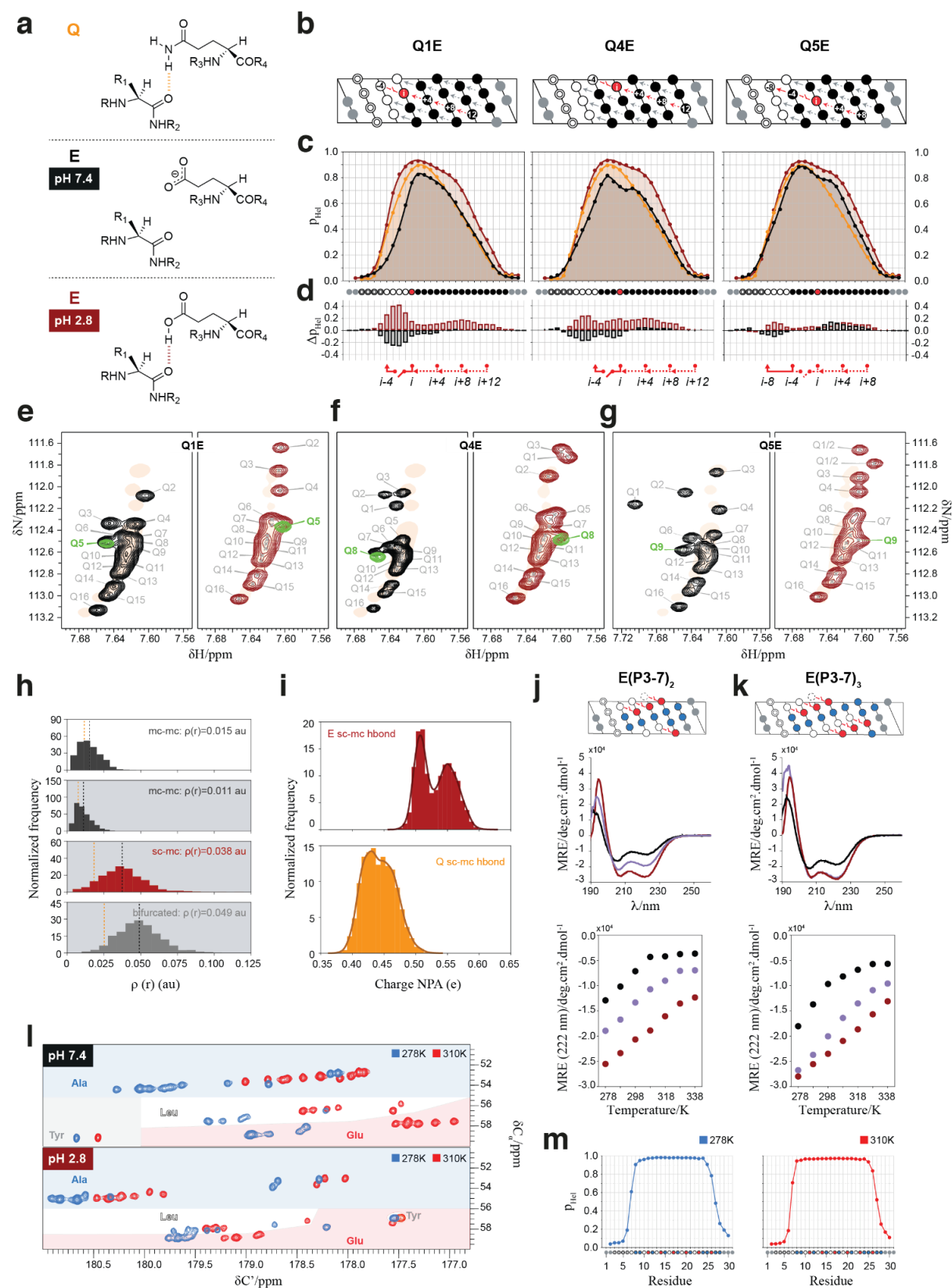
To explain the range of helicities obtained we hypothesized that it is due to two main factors: the intrinsic helical propensity of residue  $X^{30}$  and, based on our previous work<sup>11</sup>, the ability of its side chain to shield the hydrogen bond. While the former has been extensively measured<sup>23,30,32</sup>, we quantified the latter by using molecular modeling, considering that the conformation of the residue accepting the hydrogen bond ( $X_i$ ) can have an effect on the interaction of the  $\text{Gln}_{i+4} \text{H}_{\epsilon 21}$  donor with competing water molecules (Fig. 3b and S4a). For this we computed 1  $\mu\text{s}$  trajectories in different force fields<sup>29,33</sup> for all 20 possible variants in which we constrained the secondary structure of the Leu2-Gln5 segment and increased the population of the  $\text{Gln}_{i+4} \rightarrow \text{Leu}_i$  interaction with a soft restraint to facilitate sampling the relevant region of conformational space: we obtained, as expected, that the higher the frequency of the hydrogen bond, the lower the solvent accessible surface area (SASA) of the  $\text{H}_{\epsilon 21}$  atom of Gln 4, with high correlation (Fig. 3c).

Next, we quantified to what extent intrinsic helical propensity ( $x_1$ ) and  $\text{Gln}_{i+4} \text{H}_{\epsilon 21}$  SASA ( $x_2$ ) explain the experimental helical propensities (Fig. 3d) by multiple linear regression. Indeed we obtained that these two independent variables explain 73% of the variability (Fig. 3d). Remarkably, the  $\text{Gln}_4 \text{H}_{\epsilon 21}$  SASA value ( $x_2$ ) is the most important factor in the correlation, as its weight in the fitted equation is 40% higher than that of intrinsic helicity ( $x_1$ ); the model allows the prediction of the average helicities of the  $\text{L}_3\text{XQ}_{16}$  variants not included in our experimental dataset (Fig. 3e). The correlation improves (from  $r^2 = 0.73$  to 0.85) when only the subset of residue types with apolar side chains is considered, suggesting that additional factors might play a role when charged or polar side chains are present (Fig. 3f), and the results are robust to changes in MD force field<sup>29,33</sup> or intrinsic helical propensity scales (Fig. 3f). These data confirm that Leu is one of the best helicity-promoting acceptors, but that other residues such as Phe, Tyr, Ile or Met are similarly good, and that Trp is an outstanding acceptor despite its relatively low intrinsic helical propensity. Thus, residues other than Leu can be introduced as acceptors of Gln side chain to main chain interactions, increasing the versatility of our design rules.

These results prompted us to investigate the presence of  $(\text{P3-7})_n$ -like motifs in nature. To do that, we searched UniprotKB<sup>34</sup>, including the Swiss-Prot and TrEMBL databases, by using the motif search tool in ScanProsite<sup>35</sup>, which we queried with the  $\Omega\text{-X-X-(}\Omega\text{-Q-X)}_{n-1}\text{-X-Q-X-X}$  motif with  $\Omega$  denoting good acceptors of Gln side chain to main chain hydrogen bonds (namely W, L, F, Y, I, M). We found that 3451 proteins contain sequences matching the motif, belonging to organisms across the kingdoms of life with representatives of a wide variety of taxonomic lineages including archaea, bacteria, viruses and a full range of eukaryotes, from unicellular organisms to metazoa including humans (Fig. 3g). There is experimental

evidence for the existence of 94 of these proteins (UniprotKB annotation score > 3), mostly belonging to extensively characterized metazoa (Fig. S4b). Fig. 3h shows an alignment of some example sequences along with their UniprotKB accession code and the organism they belong to. We also found structural models for 50 of these sequences in the AlphaFold Database (AFDB)<sup>18</sup>: in 52% of the cases, a DSSP analysis<sup>36</sup> of the AlphaFold model shows that the (P3-7)<sub>n</sub>-like motif is helical, increasing to 76.5% when only motifs devoid of helix breaking residue types (P, G) in the central part of the sequence are considered (Fig. S4c). In summary these findings suggest that (P3-7)<sub>n</sub>-like motifs may represent a new class of naturally occurring single  $\alpha$ -helices (SAHs).

# *Glu residues as pH-sensitive conformational switches*



**Fig. 4 | Introduction of a pH-sensitive conformational switch.** **a** Schematic representation of how Gln (top) and Glu at pH 2.8 (bottom), but not Glu at pH 7.4 (center) side chains can donate a hydrogen to a main chain CO. **b** Representation as helical projection of the QxE variants of L<sub>4</sub>Q<sub>16</sub> with pH-sensitive Glu<sub>i+4</sub>→X<sub>i</sub> interactions shown as red switches and the residue substituted by E as a red circle; the color code is described in Fig. 1. **c** Residue-specific helical propensities at pH 7.4 (black) and 2.8 (dark red) compared to those of L<sub>4</sub>Q<sub>16</sub> at pH 7.4 (orange). **d** Residue-specific differences in helical propensity due to substitution of Gln by Glu at pH 7.4 (black) and pH 2.8 (dark red). **e to g** Gln side chain N<sub>ε2</sub>-H<sub>ε21</sub> regions of the 2D <sup>1</sup>H-<sup>15</sup>N HSQC NMR spectrum of QxE variants at pH 7.4 (left panel, black) and 2.8 (right panel, dark red) with, for reference, the spectrum of L<sub>4</sub>Q<sub>16</sub> as an orange shade. The spectrum of variants Q1E and Q4E with <sup>15</sup>N labeling of only the Gln in position *i*+4 to the mutated residue is superimposed in green as a verification of the assignment. **h** QM/MM-derived hydrogen bond electron densities. For reference, mean values for the Gln<sub>i+4</sub>→Leu<sub>i</sub> interaction<sup>11</sup> are shown as an orange dashed line. First and second panels: normalized histograms showing the distribution of the electron density  $\rho(r)$  of the main chain to main chain Glu<sub>i+4</sub>→Leu<sub>i</sub> hydrogen bond in the absence (white background) and in the presence (gray background) of the side chain to main chain Glu<sub>i+4</sub>→Leu<sub>i</sub> hydrogen bond. Third panel: distribution of the electron densities for the side chain to main chain Glu<sub>i+4</sub>→Leu<sub>i</sub> hydrogen bond. Fourth panel: distribution of the electron densities for the bifurcated hydrogen bond, resulting from the addition of the second and third panels. **i** Natural population analysis (NPA) charges on the hydrogen atom that can participate in the side chain to main chain hydrogen bond involving Glu (top) and Gln (down), showing that its charge depends on whether it participates in that hydrogen bond (lower values) or not. The clear peak separation in the bimodal distribution observed for Glu reports a higher electron polarization upon side chain to main chain hydrogen bonding when compared to Gln. **j to k** pH-dependent helicity and thermal stability for peptide E(P3-7)<sub>n</sub>, where all Gln residues in (P3-7)<sub>n</sub> were substituted by Glu. Top: Helical projection for the peptide with an indication of all Gln to Glu substitutions (red). Center: CD spectra (at 278 K) of peptides E(P3-7)<sub>n</sub> at pH 7.4 (black) or pH 2.8 (dark red) and of (P3-7)<sub>n</sub> at pH 7.4 (purple). Bottom: thermal denaturation of peptide E(P3-7)<sub>n</sub> (same pH and color code as above) monitored by measuring the mean residue ellipticity at 222 nm. **l** 2D CACO NMR spectra of [U-<sup>13</sup>C] labeled peptide E(P3-7)<sub>3</sub> acquired at pH 7.4 (above), pH 2.8 (below), 278 K (blue), or 310 K (red). **k** Residue-specific helical propensities for peptide E(P3-7)<sub>3</sub> at pH 2.8 and 278 K (left) or 310 K (right).

Gln to Glu substitutions in polyQ helices decrease helical character due to the inability of the Glu side chain to donate hydrogen bonds at physiological pH, where the carboxylate group is deprotonated<sup>11</sup> (Fig. 4a). Re-protonation by decreasing the pH can restore the interaction, providing us with an opportunity to introduce a pH-sensitive conformational switch in our design rules. To explore this possibility we first sought to establish whether the loss of helicity upon substitution and its restoration upon pH decrease is strictly local or whether its effects instead can propagate to other parts of the sequence due to cooperativity.

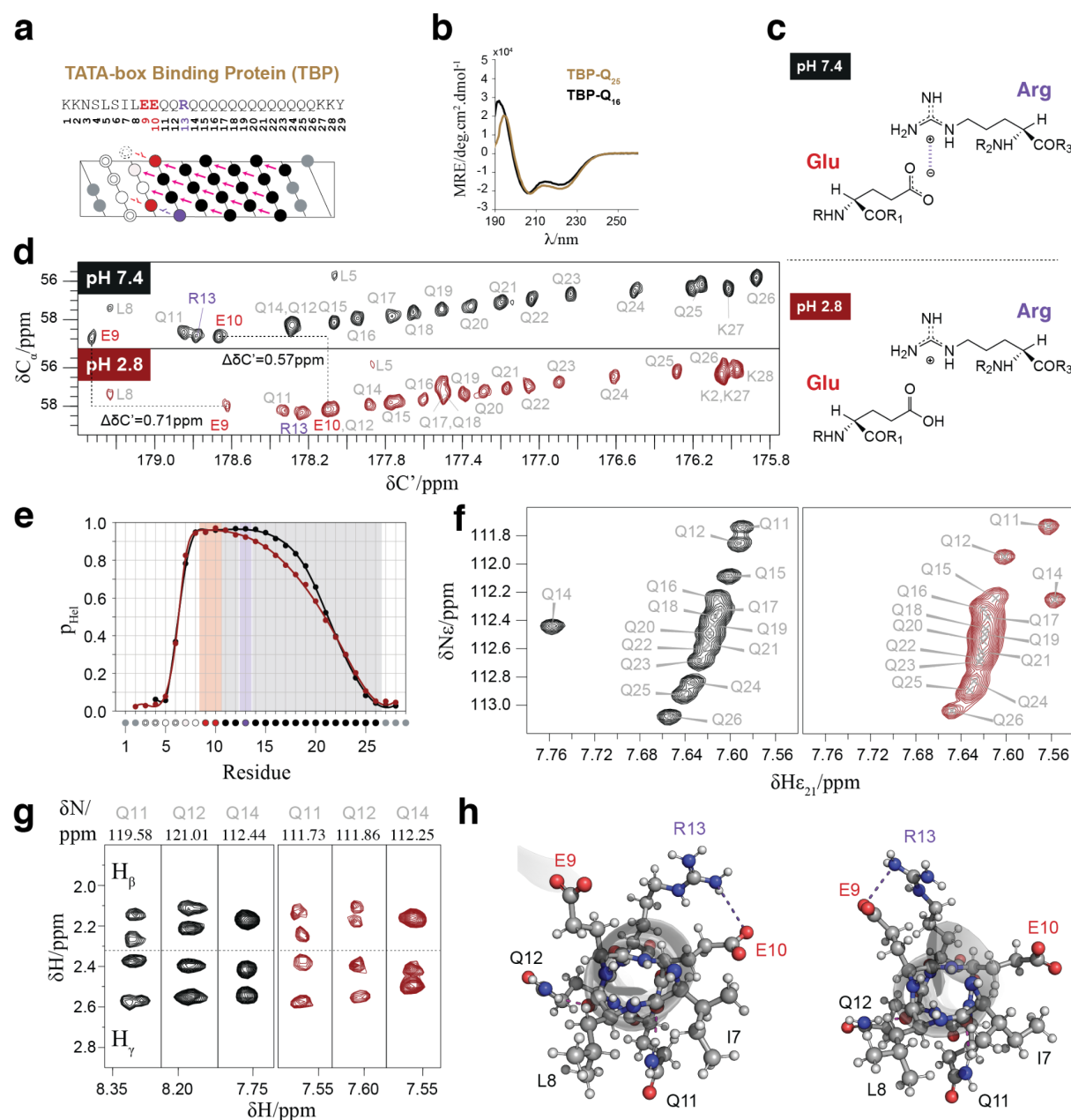
For this, we compared the residue-specific helicity of polyQ L<sub>4</sub>Q<sub>16</sub> variants with Gln to Glu substitutions at positions 1 (Q1E), 4 (Q4E), and 5 (Q5E) in the polyQ tract (Fig. 4b and S4a). In Q1E the positions experiencing the strongest decrease in helicity are Leu 2 to Leu 4, which are tethered by the first bifurcated Gln<sub>i+4</sub>→Leu<sub>i</sub> hydrogen bond in L<sub>4</sub>Q<sub>16</sub>, as expected, but there is also a small decrease in helicity for Gln residues in positions 5 to 8 (Fig. 4d, left, black). Similar effects were observed in Q4E and the loss of helical character in Q5E was

much smaller, likely because the interaction broken upon substitution, accepted by Gln<sub>i</sub>, is weak even at physiological pH. This indicates that breaking the first Gln<sub>i+4</sub>→Leu<sub>i</sub> interaction also weakens the interaction to which it is concatenated in the polyQ helix, and vice-versa, again in line with the notion that these interactions form cooperatively (Fig. 4c and d, black).

Experiments at pH 2.8, where the carboxylate group of the Glu side chain is protonated, showed that helicity was even higher than that of peptide L<sub>4</sub>Q<sub>16</sub> at physiological pH (Fig. 4c and d, red); in addition the dispersion of Gln side chain N<sub>ε</sub> and H<sub>ε21</sub> resonances in the 2D <sup>1</sup>H, <sup>15</sup>N HSQC spectrum, that is characteristic of polyQ helices, was restored (Fig 4e-g). To investigate the physical basis of this, we simulated this interaction by using a hybrid QM/MM approach (Fig. S5b and c). We found that the establishment of the side chain to main chain hydrogen bond weakened the main chain to main chain hydrogen bond: the associated average electron density decreased from 0.015 a.u. to 0.011 a.u. (Fig. 4h) and the main chain Glu<sub>i+4</sub> (H) - Leu<sub>i</sub> (O) interatomic distance increased by 0.15 Å (Fig. S5d). Consistent with the experimental stronger hydrogen bond donor character of protonated Glu, relative to that of Gln, the average electron density of the Glu<sub>i+4</sub> → Leu<sub>i</sub> side chain to main chain hydrogen bond (0.038 a.u.) was higher than that involving Gln side chains (0.017 a.u.) and the electronic polarization of the H<sub>ε2</sub> donor was stronger for Glu relative to Gln (Fig. 4i).

Finally, to test whether switchable side chain to main chain hydrogen bonds can be integrated in our design rules, we studied variants of the (P3-7)<sub>2</sub> and (P3-7)<sub>3</sub> peptides where all Gln residues were substituted by Glu, namely E(P3-7)<sub>2</sub> and E(P3-7)<sub>3</sub> (Figs. 4j and 4k, top). As expected, these were less helical than (P3-7)<sub>2</sub> and (P3-7)<sub>3</sub> at physiological pH, but more at pH 2.8 (Figs. 4j and 4k, center). Both the switchable nature and the increased strength of the side chain to main chain hydrogen bonds involving Gln is apparent in the enhanced thermal stability of the E(P3-7)<sub>n</sub> variants (Figs. 4j and 4k, bottom): while at physiological pH peptide E(P3-7)<sub>3</sub> loses its helical character at a lower temperature than (P3-7)<sub>3</sub>, at pH 2.8 it remains helical even at the highest tested temperature, 340 K. Indeed <sup>13</sup>C-detected 2D CACO NMR spectra confirmed that, at pH 2.8, E(P3-7)<sub>3</sub> was highly helical at 310 K, the physiological temperature (Figs. 3l and m).

$Gln_{i+4} \rightarrow X_i$  interactions can be combined with electrostatic interactions between side chains



**Fig. 5 | The polyQ tract of TBP forms a helix stabilized by Gln side chain to main chain hydrogen bonds and an electrostatic interaction . a** Sequence, numbering and helical projection of peptide TBP-Q<sub>16</sub>. The electrostatic interaction is color-coded in purple. **b** CD spectra of peptides TBP-Q<sub>16</sub> (black) and TBP-Q<sub>25</sub> (gold). **c** Scheme of the electrostatic interaction between Arg and Glu at physiological pH and its absence at acidic pH. **d** <sup>13</sup>C-detected CACO spectra of the TBP polyQ tract at pH 7.4 (top, black) and 2.8 (dark red, down). **e** Residue-specific helicity of TBP at pH 7.4 (black) and 2.8 (dark red). Shades color-coded as in **a** are shown to guide the eye. **f** Region of the <sup>1</sup>H-<sup>15</sup>N HSQC spectra of TBP at pH 7.4 (left, black) and 2.8 (right, dark red) showing the N<sub>ε2</sub>-H<sub>ε21</sub> correlations. **g** Strips from the 3D H(CC)(CO)NH spectra of TBP at pH 7.4 (left, black) and 2.8 (right, dark red) displaying the side chain aliphatic <sup>1</sup>H resonances of the first three Q residues in the polyQ tract. Strips were chosen from either N<sup>H</sup> or N<sub>ε2</sub> for clarity. **h** Frames from the MD trajectory obtained for TBP. An

orthogonal view of the helix is shown. In the left panel, a frame is shown where Arg13 (*i*) establishes an electrostatic interaction with E10 (*i*-3), represented by the purple dashed line. Simultaneously, both Gln11 and Gln12 establish bifurcated hydrogen bonds with Ile7 and Leu8, respectively (pink dashed lines). Right: a frame where Arg13 (*i*) establishes an electrostatic interaction with Glu9 (*i*-4), while the two bifurcated hydrogen bonds previously described are also present.

The natural single  $\alpha$ -helices studied until now are stabilized by numerous electrostatic interactions between side chains of opposite charge at relative positions  $i, i+3$  or  $4$ <sup>37</sup>. We sought to investigate whether they can be combined with  $\text{Gln}_{i+4} \rightarrow \text{X}_i$  interactions to stabilize  $\alpha$ -helices. For this we studied the polyQ tract of the TATA-box binding protein (TBP), that has a primary structure that suggests the presence of an electrostatic interaction, between either of two Glu residues immediately flanking the tract at the N-terminus (Glu9 and Glu10) and an Arg interrupting it (Arg13) (Fig. 5a). This interaction can occur concomitantly with two strong bifurcated hydrogen bonds accepted by Ile7 and Leu8, at position *i*-4 relative to the first two Gln residues of the tract. As observed for the polyQ tracts in AR<sup>11</sup> and huntingtin<sup>38</sup> the CD spectrum of a peptide spanning a tract of size 16 and its N-terminal flanking region, TBP-Q<sub>16</sub>, showed it is strongly helical and that its expansion to 25 Gln residues increases its helicity (Fig. 5a,b and S6).

We then used NMR to characterize TBP-Q<sub>16</sub> at residue resolution (Fig. 5d,e). At physiological pH, in agreement with the CD data, the peptide forms a fully folded helix between residues Glu9 and Gln14 and its helicity decreases progressively towards the C-terminus; at acidic pH, instead (Fig. 5c), at which Glu side chains are protonated, the helical propensity starts decreasing at position 12. In addition, both the spectral signature of concatenated  $\text{Gln}_{i+4} \rightarrow \text{X}_i$  interactions (Fig. 5f) and the rotamer selection associated with these interactions (Fig. 5g) are diminished for Gln14 ( $H_\gamma$ ) and, to some extent, Gln12 ( $H_\beta$ ). These results are in agreement with the formation of a helix-stabilizing electrostatic interaction between Glu9 (or Glu10) and Arg13 that is lost upon protonation at low pH, confirming that electrostatic interactions can be combined with Gln side chain to main chain interactions. To further confirm that these interactions can co-exist we simulated a WTE-enhanced MD trajectory of TBP-Q<sub>16</sub><sup>29,39</sup>: Fig. 5h shows two frames of the trajectory in which both the Gln11  $\rightarrow$  Ile7 and the Gln12  $\rightarrow$  Leu8 bifurcated hydrogen bonds occur simultaneously with a salt bridge involving Arg13 and either Glu10 (left) or Glu9 (right).

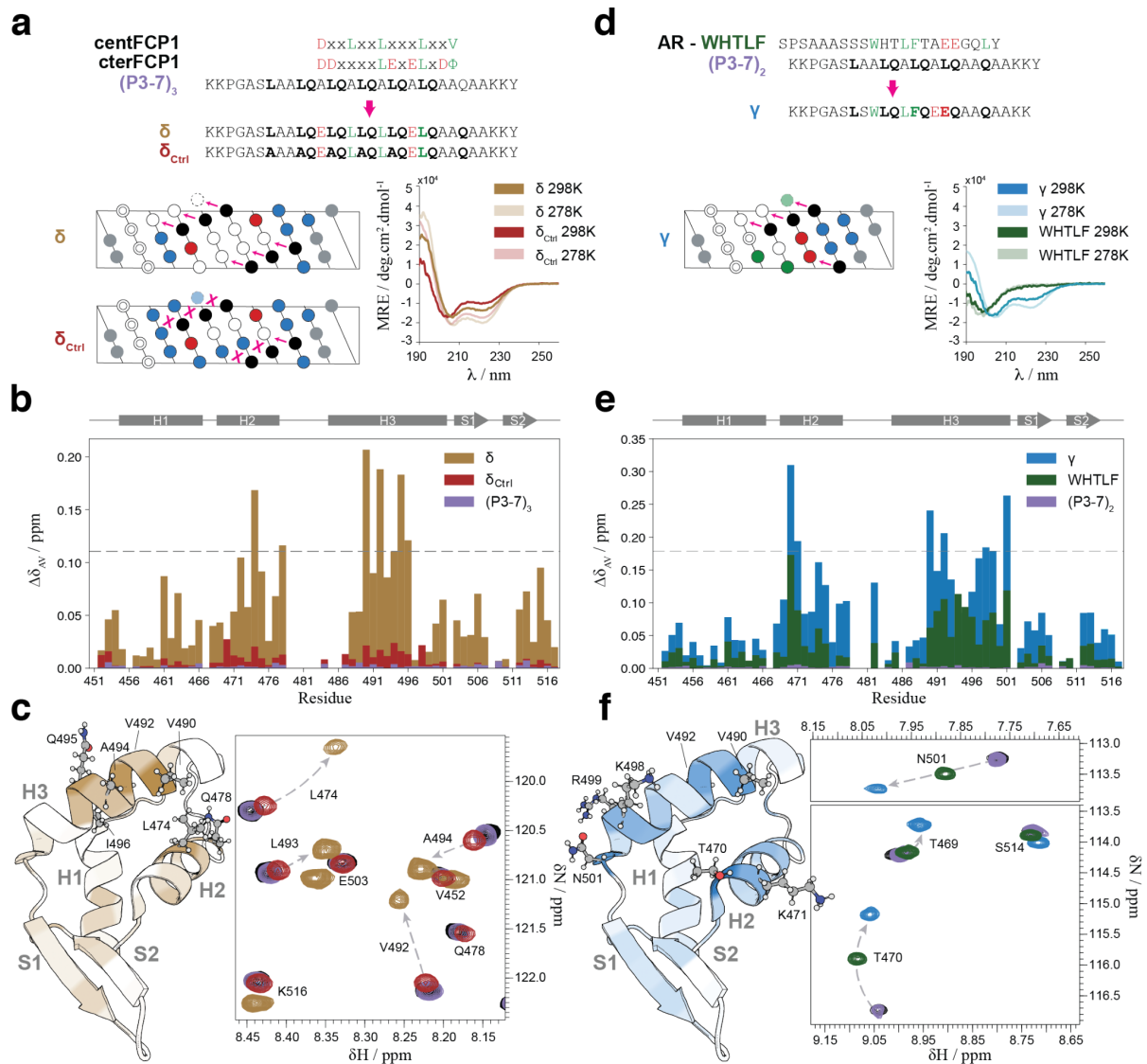
### *Design of Gln-based single $\alpha$ -helices that bind a globular target*

Gln-based  $\alpha$ -helical peptides can be modified to bind specific globular targets: the Ala residues in the (P3-7)<sub>n</sub> scaffolds (Fig. 1) can indeed be modified at will because they are not involved in the interactions that stabilize the helical structure. To prove this concept we

modified the sequence of peptide (P3-7)<sub>3</sub> to interact with the C-terminal domain of RAP74 (RAP74-CTD), a small globular domain that binds to intrinsically disordered motifs that fold upon binding<sup>40–42</sup>. We blended the sequence of (P3-7)<sub>3</sub> with that of two different motifs (centFCP1 and cterFCP1) derived from FCP1 that interact with this globular protein independently<sup>42</sup>. This led to peptides  $\delta$  and  $\delta_{\text{ctrl}}$ :  $\delta$  was designed to bind to RAP74-CTD whereas  $\delta_{\text{ctrl}}$  is a control sequence equivalent to  $\delta$  where we replaced Leu by Ala that, despite having high helical propensity, are bad acceptors of side chain to main interactions, thus destabilizing helicity (Fig. 6a and S7).

An analysis of the structural properties of peptides  $\delta$  and  $\delta_{\text{ctrl}}$  by CD showed, as expected, that the former is more helical than the latter, especially at room temperature (298 K), further confirming the important role of Leu residues for the stability of the helical fold (Fig. 6a). We then analyzed the chemical shift perturbations (CSPs) in the <sup>1</sup>H, <sup>15</sup>N BEST-TROSY spectrum of RAP74-CTD induced by peptide binding. We obtained that  $\delta$  induced perturbations in residues of the globular protein that define the binding site of FCP1 in this globular target<sup>42</sup>, confirming a similar binding mode. This was in contrast to the results obtained with both  $\delta_{\text{ctrl}}$  and (P3-7)<sub>3</sub>, which in both cases failed to interact (Fig. 6b and c). The result obtained with  $\delta_{\text{ctrl}}$  indicates that the helical character of  $\delta$  is key for its ability to interact with RAP74-CTD whereas that obtained with (P3-7)<sub>3</sub> indicates that helical character does not suffice, and that the identity of the residues placed in the vacant positions is indeed key for binding, in agreement with our hypothesis.

To provide a second proof of concept we blended the sequence of (P3-7)<sub>2</sub> with that of a motif found in the activation domain of AR that also interacts with RAP74-CTD, to yield peptide  $\gamma$  (Fig. 6d). As we previously showed, inhibiting this interaction with small molecules or peptides is a potential avenue to treat castration-resistant prostate cancer<sup>40</sup>. Linear peptides spanning the AR motif bind weakly to RAP74-CTD due, at least in part, to their low helical propensity, providing us with an additional opportunity to test the potential of our designs. In peptide  $\gamma$  two Gln<sub>i+4</sub>→X<sub>i</sub> hydrogen bond acceptor positions were modified to accommodate the binding motif following the rules learned previously (Fig. 3). In agreement with our hypothesis we obtained that peptide  $\gamma$  binds more strongly than either (P3-7)<sub>2</sub> or the WHTLF peptide (Fig. 6e and f). In summary we have shown that the Gln-based helical peptides can be adapted to interact with specific globular proteins.



**Fig. 6 | Design of Gln-based single helical peptides that interact with RAP74-CTD** **a** Top: design of peptide  $\delta$ , blending the consensus RAP74-CTD binding motifs of FCP1 with peptide (P3-7)<sub>3</sub>. Peptide  $\delta_{ctrl}$ , in which Leu residues in  $\delta$  are substituted by Ala, is also shown. Bottom, left: projections of the fully helical structures of peptides  $\delta$  and  $\delta_{ctrl}$ . Bottom, right: CD spectra of peptides  $\delta$  (gold) and  $\delta_{ctrl}$  (dark red) at 298 K (saturated) and 278 K (pastel). **b** Averaged chemical shift perturbations observed in the <sup>1</sup>H-<sup>15</sup>N resonances of the RAP74-CTD upon the addition of 15 molar equivalents of the (P3-7)<sub>3</sub> (purple), the  $\delta$  (gold) and the  $\delta_{ctrl}$  (dark red) peptides. The dashed line sets the threshold for the top 10% peaks with most intense averaged CSPs. **c** Left: structure of the RAP74-CTD domain of TFIIF (PDB entry 1NHA) colored in a gradient representing the averaged CSPs observed upon the addition of 15 molar equivalents of the  $\delta$  peptide (white, minimum CSP<sub>av</sub>; gold, maximum CSP<sub>av</sub>). Right: region of the overlaid <sup>1</sup>H-<sup>15</sup>N BEST-TROSY spectra of RAP74-CTD in the absence (black) and in the presence of 15 molar equivalents of the (P3-7)<sub>3</sub> (purple), the  $\delta$  (gold) and the  $\delta_{ctrl}$  (dark red) peptides. **d** Top: design of peptide  $\gamma$  blending the RAP74-CTD binding motif AR-WHTLF with peptide (P3-7)<sub>2</sub>. Bottom, left: projection of the fully helical structure of peptide  $\gamma$ . Bottom, right: CD spectra of peptide  $\gamma$  (blue) and AR-WHTLF (green) at 298 K (saturated) and 278 K (pastel). **e** as in **b** for (P3-7)<sub>2</sub> (purple), peptide  $\gamma$  (blue) and peptide AR-WHTLF (green). **f**

Left: as in **c**, left, for peptide  $\gamma$ . Right: as in **c**, right, for the (P3-7)<sub>2</sub> (purple), the  $\gamma$  (blue) and the AR-WHTLF (green) peptides.

## Discussion

### *A new tool to design fully helical peptides*

Our results show that Gln side chain to main chain hydrogen bonds can be used to design linear peptides that fold into  $\alpha$ -helices (Fig. 2) with properties that make them attractive for various applications: they are highly soluble, even upon thermal denaturation, are not stabilized by electrostatic interactions, unlike the Glu and Lys/Arg-rich single  $\alpha$ -helices reported until now<sup>43–45</sup>; and display some degree of folding cooperativity due to the concatenation of side chain to main chain hydrogen bonds explicit in our design rules (Fig. 1). Our data shed light on the potential basis of such cooperative effect, which occurs when the donor of a  $\text{Gln}_{i+4} \rightarrow \text{Leu}_i$  interaction and the acceptor of the next one share a peptide bond.

An important feature of our design rules is their versatility: the residue accepting the hydrogen bond donated by the Gln side chain can be any residue able to shield the interaction from competition with water, such as Trp, Leu, Phe, Tyr, Met and Ile (Fig. 3). Remarkably, we find natural sequences fulfilling our design in different kingdoms of life suggesting that they represent a new class of uncharged single  $\alpha$ -helices (SAHs) that had remained undetected<sup>43–45</sup>. In addition, in suitable cases the design can be complemented by electrostatic interactions between side chains of opposite charge and Gln residues can be mutated to Glu to introduce a pH-responsive conformational switch that uses only natural amino acids<sup>46</sup> and does not involve changes in oligomerization state<sup>47</sup> (Figs. 4 and 5).

The key feature of our scaffold design is that it defines the identity of just a fraction of the peptide residues: the rest can be chosen or optimized for specific applications. To prove this concept, we designed two peptides to bind the globular target RAP74-CTD by using an approach analogous to previous motif-grafting attempts on folded scaffolds<sup>48</sup> (Fig. 6). In this specific case, naively blending these sequence features with the sequence of the designed (P3-7)<sub>n</sub> scaffold proved sufficient for successful targeting. The affinities that we have obtained in this initial exercise (Fig. S6d and e) can be sufficient for certain applications<sup>49</sup> and we anticipate that it will be possible to greatly improve them by systematically searching the sequence space available using techniques for affinity maturation based on high-throughput mutational scans<sup>50</sup>, especially when taking advantage of the versatility of our design rules.

### *A rationale for the structural heterogeneity of polyQ tracts*

The link between polyQ tract expansion and disease onset in polyQ disorders has not been established<sup>51</sup>. These tracts are found in intrinsically disordered regions, and much work has been devoted to investigating whether expansion changes their structural properties. The results obtained have been inconclusive: single molecule Förster resonance energy transfer (FRET) and NMR measurements showed little influence of tract length on the conformation of the tract found in huntingtin<sup>13,52</sup> whereas recent studies from some of us have instead shown that the helical propensity of the tract found in the androgen receptor (AR) increases upon expansion<sup>11,12</sup>, as does the tract found in TBP (Fig. 5b).

Our results help rationalize these observations by considering that polyQ tracts are in a polyQ helix-coil equilibrium. Its position can be influenced by the residues flanking the tract at its C-terminus<sup>13,53</sup> but, for a given set of solution conditions and tract length, it is mainly determined by the four residues flanking the tract at its N-terminus (Fig. 3). When these are good acceptors of Gln side chain to main chain hydrogen bonds, as in AR (LLLL), the polyQ helix is favored. Instead, when only two are good acceptors as in huntingtin (LKSF), the coil is favored<sup>13,19,52,54</sup>. The results obtained for TBP also fit this working model: its flanking region (ILEE) contains two good acceptors (Ile, Leu) and the two other residues (Glu) can establish, at physiological pH, an electrostatic interaction with the Arg residue three or four positions towards the C-terminus, favoring the polyQ helix.

The polyQ helix-coil equilibrium is sensitive to solution conditions: the entropic cost of folding results in higher stability of polyQ helices at relatively low temperatures whereas high temperatures favor the coil. This contributes in part to explaining the discrepancy between the results obtained with huntingtin, carried out at room temperature<sup>52</sup>, and those obtained for the androgen receptor, carried out at 5°C<sup>11</sup>: indeed, CD studies of the structural properties of huntingtin showed an increase in helical propensity upon tract expansion at low temperature (-10°C)<sup>38</sup>. Finally our observation that pairs of concatenated side chain to main chain interactions form cooperatively contributes to explaining how expansion shifts the equilibrium to the polyQ helix state both in AR<sup>11</sup> and in huntingtin<sup>52</sup>.

In summary our results suggest that polyQ tracts are in a helix-coil equilibrium that is governed by the N-terminal flanking region, by solution conditions and, due to cooperativity, by tract length. Given that interactions between low-populated helical conformations of huntingtin play a role in the early stages of its aggregation into amyloid fibrils<sup>55,56</sup> we

speculate that polyQ expansion leads to the onset of Huntington's disease at least in part by stabilizing pre-nucleation oligomeric species that are on-pathway to aggregation.

### *Concluding remarks*

In summary we have shown how an appropriate concatenation of Gln side chain to main chain hydrogen bonds makes it possible, on the one hand, to design highly helical peptides that can be tailored to specific applications and, on the other hand, to rationalize the until now perplexing observations regarding the structural properties of polyQ tracts. We anticipate that the knowledge gained about this interaction will influence future developments in peptide design, particularly in the use of peptides as therapeutics, as well as contribute to better understanding the molecular basis of polyQ diseases.

## Methods

### *Peptide sample preparation*

Recombinant peptides with  $^{15}\text{N}$  or  $^{13}\text{C}$ ,  $^{15}\text{N}$  isotope enrichment were prepared as detailed elsewhere<sup>57</sup>. Synthetic genes coding for the peptide of interest and codon-optimized for expression in *Escherichia coli*, with an N-terminal His<sub>6</sub>-Sumo tag fusion and cloned into the pDEST-17 vector, were directly obtained from GeneArt (Thermo Fisher Scientific, Waltham, MA, USA). Genes coding for the L<sub>4</sub>Q<sub>16</sub> variants in the L<sub>3</sub>X and QXE series (see Table S1) were obtained using the Q5 Site-Directed Mutagenesis Kit from New England Biolabs (Ipswich, MA, USA). Rosetta (DE3)pLysS competent cells (Novagen, Merck KGaA, Darmstadt, Germany) were used for expression in M9 medium containing  $^{15}\text{NH}_4\text{Cl}$  and, where required,  $^{13}\text{C}$ -glucose (both from Cambridge Isotope Laboratories Inc., Tewksbury, MA, USA) as the sole nitrogen and carbon sources, respectively. All purification steps were performed at 277 K. Cell lysates in lysis buffer (20 mM Tris-HCl, 100 mM NaCl, 20 mM imidazole, pH 8.0) were purified by immobilized metal affinity chromatography (IMAC) using a HisTrap HP 5 mL column mounted on an Äkta Purifier System (GE Healthcare, Chicago, IL, USA). Fractions containing the His<sub>6</sub>-Sumo-peptide fusion in elution buffer (20 mM Tris-HCl, 100 mM NaCl, 500 mM imidazole, pH 8.0) were pooled and dialyzed overnight in lysis buffer while treated with ubiquitin-like specific protease Ulp1 (50 µg/mL). The His<sub>6</sub>-Sumo tag was removed with an additional IMAC step and the peptide-containing flow-through was dialyzed in ultrapure MilliQ water before lyophilization. Unlabeled synthetic peptides from solid-phase peptide synthesis were directly obtained from Genscript (Piscataway, NJ, USA) as lyophilized powder with >95% purity. Both recombinant and synthetic lyophilized peptides were dissolved in 6M guanidine thiocyanate (Merck KGaA, Darmstadt, Germany) and incubated overnight at 1250 rpm, 298 K, in a thermoblock. The sample was then injected in an Äkta Purifier System equipped with a Superdex Peptide 10/300 GL column (GE Healthcare, Chicago, IL, USA) equilibrated with ultrapure water, 0.1% trifluoroacetic acid. The fractions containing monomeric peptide were pooled and centrifuged at 386000 g for 3 h in an Optima TLX ultracentrifuge equipped with a TLA 120.1 rotor (Beckman Coulter, Atlanta, GA, USA). Orthophosphoric acid or sodium phosphate were added to a final concentration of 20 mM to adjust the pH to 2.8 or 7.4, respectively. The peptide concentration was determined by measuring the absorbance at 280 nm (extinction coefficients were calculated using the ProtParam tool on the ExPASy website, <https://web.expasy.org/protparam>) before diluting the sample to the final experimental concentration.

### *Protein sample preparation*

Recombinant samples with  $^{15}\text{N}$  isotope enrichment of the C-terminal domain of subunit 1 of the general transcription regulator TFIIF (RAP74-CTD), spanning residues 450-517, were obtained as described previously<sup>40</sup>. A codon-optimized synthetic gene cloned in a pDONR221 vector was obtained from GeneArt (Thermo Fisher Scientific, Waltham, MA, USA) and subcloned into a pDEST-His<sub>6</sub>MBP vector obtained from Addgene. Rosetta (DE3)pLysS competent cells (Novagen, Merck KGaA, Darmstadt, Germany) were grown in MOPS medium with  $^{15}\text{NH}_4\text{Cl}$  as the only nitrogen source (310 K, induction at  $\text{OD}_{600} = 0.7$ , 1 mM IPTG, harvesting 3 h after induction). Soluble fractions of cell lysates from sonication in lysis buffer (50 mM Tris-HCl, 1M NaCl, 10 mM imidazole, pH 8.0) were purified by IMAC and fractions containing the His<sub>6</sub>MBP-TEV-RAP74-CTD fusion were dialyzed at 277 K overnight against cleavage buffer (50 mM Tris-HCl, 200 mM NaCl, 0.5 mM EDTA, pH 8.0) in the presence of TEV protease (50  $\mu\text{g}/\text{mL}$ ). A second IMAC step was performed and the flow-through containing the RAP-CTD was loaded onto a HiTrap SP HP cation exchange column followed by a size exclusion chromatography step on a Superdex 75 GL 10/300 column equilibrated with NMR buffer (20 mM sodium phosphate, 0.1% trifluoroacetic acid, pH 7.4), both mounted in an Äkta Purifier System (GE Healthcare, Chicago, IL, USA). The sample was concentrated to 100  $\mu\text{M}$  using an Amicon Ultra 15 mL centrifugal filter (Merck KGaA, Darmstadt, Germany).

### *CD spectroscopy*

Peptide samples for CD spectroscopy were diluted to a final concentration of 30  $\mu\text{M}$  in a volume of 400  $\mu\text{L}$  in either 20 mM phosphoric acid (pH 2.8) or sodium phosphate (pH 7.4) buffer. Spectra were obtained at 278 K (unless stated otherwise) in a Jasco 815 UV spectro-photopolarimeter with a 1 mm optical path cuvette using a data interval of 0.2 nm in the 190 - 260 nm range with a scanning speed of 50 nm  $\text{min}^{-1}$  and 20 accumulations. A blank spectrum acquired on the pertaining buffer under the same experimental conditions was subtracted from the sample spectrum. Thermal denaturation experiments were performed by acquiring a single accumulation spectrum with the same parameters at 10 K intervals, with a temperature ramp speed of 10 K  $\text{min}^{-1}$  and an equilibration time of 1 min.

### *NMR spectroscopy*

Peptide samples for NMR spectroscopy were diluted to a final concentration of 100  $\mu\text{M}$  in a volume of 400  $\mu\text{L}$  in either 20 mM phosphoric acid (pH 2.8) or sodium phosphate (pH 7.4) buffer with added 10% v/v  $\text{D}_2\text{O}$  and 10  $\mu\text{M}$  DSS for internal chemical shift referencing, then

filled into Shigemi tubes (Shigemi Co. Ltd, Tokyo, Japan). All NMR experiments were recorded at 278 K (unless stated otherwise) on either a Bruker Avance III 600 MHz or a Bruker Avance NEO 800 MHz spectrometer, both equipped with TCI cryoprobes. Unlabeled synthetic peptides P1-5, P2-6, P3-7, P5-9 were characterized by two-dimensional homonuclear (TOCSY and NOESY) and heteronuclear ( $^1\text{H}$ - $^{13}\text{C}$  HSQC) experiments. The TOCSY and NOESY mixing times were set to 70 and 200 ms, respectively. Water suppression was achieved by excitation sculpting<sup>58</sup> using a 2 ms long Squa100.1000 selective pulse. For peptide backbone resonance assignment, using uniformly  $^{15}\text{N}$ ,  $^{13}\text{C}$  labeled peptides, the following series of 3D triple resonance BEST-TROSY<sup>59</sup> experiments were acquired with 25% non-uniform sampling (NUS): HNCO, HN(CA)CO, HN(CO)CA, HNCA, and HN(CO)CACB. For some peptides we resolved assignment ambiguities by also acquiring a 3D (H)N(CA)NH spectrum. Furthermore, 2D  $^{13}\text{C}$ -detected CACO and CON experiments<sup>60</sup> were measured. Data processing was carried out with qMDD<sup>61</sup> for non-uniform sampled data and with NMRPipe<sup>62</sup> for all uniformly collected experiments. Data analysis was performed with CcpNmr Analysis<sup>63</sup>. Determination of the residue-specific helical propensity ( $p_{\text{Hel}}$ ) from backbone chemical shifts  $\text{H}^{\text{N}}$ ,  $\text{N}^{\text{H}}$ ,  $\text{C}'$  and  $\text{C}_{\alpha}$  was performed using CheSPI<sup>14</sup>, which uses sequence and condition-corrected (temperature, pH) estimates for the reference random coil chemical shifts derived from POTENCI<sup>15</sup>. CheSPI was chosen over other algorithms as it intrinsically considers chemical shift changes derived from Glu side chain protonation, as the  $\sim 0.6$  ppm  $\text{C}'$  chemical shift change reported before<sup>64</sup> and observed in the TBP CACO spectra (Fig. 5d).

Side chain aliphatic  $^1\text{H}$  chemical shifts were obtained from 3D  $^{15}\text{N}$  edited TOCSY-HSQC (75 ms mixing time) and NOESY-HSQC (200 ms mixing time) spectra. Glutamine side chain resonances were assigned using complementary 3D H(CC)(CO)NH and (H)CC(CO)NH spectra recorded with 25% NUS and 14 ms C,C-TOCSY mixing. To further confirm the side chain  $\text{N}_{\epsilon}$  assignments for Gln5 in peptide Q1E and Gln8 in peptide Q4E we recorded 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectra of synthetic unlabeled peptides with specific  $^{15}\text{N}$  labeling only in the positions of interest.

For the detailed analysis of sidechain rotamer distributions by the CoMAND approach, a 3D CNH-NOESY (i.e. 3D [H]C,NH HSQC-NOESY-HSQC) spectrum<sup>65</sup> of [ $\text{U}-^{13}\text{C}$ ,  $^{15}\text{N}$ ] labeled (P3-7)<sub>2</sub> was recorded at 800 MHz, 278 K, with 400 ms NOE mixing time and  $64(^{15}\text{N}) \times 86(^{13}\text{C}) \times 2048(^1\text{H})$  complex data points corresponding to  $15.2 \times 104.8 \times 6.3$  Hz FID resolution. The final  $^{15}\text{N}$ -HSQC module employed sensitivity-enhanced coherence selection by gradients and band-selective flip-back of  $\text{H}^{\text{C}}$  polarization to enable its faster re-equilibration during a shorter total interscan delay (experimentally optimized as 0.6 s). For

prior assignment of all aliphatic side chain  $^1\text{H}$  and  $^{13}\text{C}$  resonances and easy distinction between intra- and inter-residual NOE signals in the  $[\text{H}]\text{C},\text{NH}$  HSQC-NOESY-HSQC spectrum, we furthermore recorded a set of 3D  $[\text{H}]\text{CC}[\text{CA}]\text{NH}$  TOCSY (11.3 and 22.6 ms FLOPSY8 mixing) and  $[\text{H}]\text{CC}[\text{CO}]\text{NH}$  TOCSY spectra (9 and 18 ms FLOPSY8 mixing).

To study glutamine side chain  $^{15}\text{N}$  relaxation, a  $^{15}\text{N}$  labeled  $(\text{P3-7})_2$  sample was prepared in NMR buffer (pH 7.4). To avoid bias due to dipole-dipole cross-correlated relaxation within the  $\text{N}_{\epsilon 2}\text{H}_2$  moieties and thus allow a direct comparison with the main chain NH data, we sampled only their 50% semi-protonated  $\text{N}_{\epsilon 2}\text{HD}$  isotopomers in buffered 50%  $\text{D}_2\text{O}$  and applied continuous deuterium decoupling during the  $^{15}\text{N}$  coherence evolution. Of note, the differential deuterium isotope shift of  $^{15}\text{N}_{\epsilon 2}$  in the  $\text{N}_{\epsilon 2}\text{H}_{\epsilon 21}\text{D}_{\epsilon 22}$  vs  $\text{N}_{\epsilon 2}\text{D}_{\epsilon 21}\text{H}_{\epsilon 22}$  species<sup>66</sup> also allowed an unambiguous stereospecific signal assignment of the attached side chain carboxamide  $\text{H}_{\epsilon}$  (Fig. S2b). To measure  $^{15}\text{N}$   $R_1$  and  $R_2$  rates the conventional pulse sequences with sensitivity-enhanced coherence selection by gradients, water flip-back, and fully interleaved acquisition of relaxation delays were complemented with continuous deuterium decoupling during  $t_1(^{15}\text{N})$  in order to suppress  $^{15}\text{N}(t_1)$  line broadening from scalar relaxation (via  $^1\text{J}_{^{15}\text{N},\text{D}}$  coupling) for the glutamine sidechain NHD isotopomers of interest. In contrast, the pulse sequence for measuring the  $^{15}\text{N}\{^1\text{H}\}$  heteronuclear NOE (likewise with sensitivity-enhanced coherence selection by gradients and continuous deuterium decoupling during  $t_1(^{15}\text{N})$ ) required further critical adaptations to suppress detrimental antiphase signal components (in  $\text{F1}(^{15}\text{N})$ ) for the 50% glutamine  $\text{NH}_2$  isotopomers that impede a clean quantification of nearby NHD signals of interest. Thus, for the reference (non-saturated) spectrum, the first  $90^\circ$   $^1\text{H}$  pulse in the sensitivity-enhanced reINEPT following  $t_1(^{15}\text{N})$  had to be cycled (inverted) along with the receiver phase. For the  $\text{H}^{\text{N}}$  saturated spectrum, however, further antiphase contamination derives from some  $4\text{N}_2\text{H}^{\pm}\text{H}^{\pm}$  multiquantum coherence forming during the  $\text{H}^{\text{N}}$  saturation sequence<sup>67</sup> that can be removed by its phase cycling and/or by appending a concatenated  $^1\text{H}$  spoil sequence (z-gradient 1 -  $90^\circ(^1\text{H})$  - z-gradient 2).

Residual dipolar couplings (RDCs) were determined from the difference between couplings observed for aligned versus unaligned  $(\text{P3-7})_2$  samples (with  $\text{U-}^{15}\text{N},^{13}\text{C}$  labeling) where alignment was achieved<sup>68</sup> using a gel kit from New Era Enterprises, Inc. (Vineland, NJ, USA). For this, 7% acrylamide gels were dialyzed in ultrapure water for 3 h and NMR buffer overnight. The prepared gels were then soaked in the peptide sample (ca. 0.2 mM) overnight at 277 K and squeezed into open-ended 5 mm NMR tubes using a funnel and piston. The filled tube was closed with a bottom plug and a Shigemi top plunge. The sample alignment uniformity was assessed via the deuterium signal splitting. One bond  $^1\text{H-}^{15}\text{N}$  RDCs were obtained from comparing aligned versus unaligned 2D BEST-TROSY spectra<sup>59</sup> measured at

278 K and selecting either the  $H^N$  TROSY or semi-TROSY signals in the direct dimension. One bond  $^{13}C'-^{15}N$  and two bond  $^{13}C'-^1H^N$  RDCs were derived by comparing the pertaining  $^1J_{C',N}$  splitting in the indirect ( $^{15}N$ ) and  $^2J_{C',H}$  splitting in the direct ( $H^N$ ) dimension, respectively, observed in the (not  $^{13}C$  decoupled) 2D  $^1H,^{15}N$  HSQC spectra of aligned vs unaligned samples measured at 278 K. PALES<sup>69</sup> was used to calculate the expected RDCs for each structure in the CoMAND-derived ensemble, allowing for the calculation of an independent alignment tensor for each frame based only on its coordinates. Predicted RDCs were obtained as the average of the ca. 200 frames generated in the 20 iterations of converged, R-factor-minimizing CoMAND global ensemble calculations.

To study molecular recognition we measured the 2D  $^1H-^{15}N$  BEST-TROSY spectrum of  $^{15}N$ -labeled Rap74-CTD (50  $\mu M$  throughout, 298 K) in the presence of increasing concentrations of peptide. Before the experiment, both the protein and the peptide were dialyzed (277 K, two dialysis steps) in the same preparation of NMR buffer (20 mM sodium phosphate, 0.1% trifluoroacetic acid, pH 7.4), using a Pur-A-Lyzer (Sigma-Aldrich, Burlington, MA, USA) and a Micro Float-A-Lyzer (Spectrum Laboratories, San Francisco, CA, USA) respectively. The chemical shift assignment of RAP74-CTD was reported previously (BMRB code 27288). Averaged  $^1H$  and  $^{15}N$  chemical shift perturbations (CSPs) were calculated as:

$$\Delta\delta_{AV} = \sqrt{(\Delta\delta_{1H})^2 + \left(\frac{1}{5}\Delta\delta_{15N}\right)^2}$$

Dissociation constants,  $K_D$ , and averaged CSP amplitude,  $\Delta\delta_{max}$ , for  $\delta$  and  $\gamma$  peptide binding to Rap74-CTD were obtained by a global fitting (nonlinear regression) of the 10% peaks with the largest averaged CSPs to the following single-site binding model (1:1 stoichiometry):

$$\Delta\delta = \Delta\delta_{max} \frac{([Protein] + [Peptide] + K_D) - \sqrt{([Protein] + [Peptide] + K_D)^2 - 4[Protein][Peptide]}}{2[Protein]}$$

A Monte Carlo simulation varying both protein and peptide concentrations within 20% experimental errors was used to derive error margins for the final  $K_D$  values.

### *Structural characterization using CoMAND*

To investigate the conformational tendencies of the (P3-7)<sub>2</sub> peptide we applied the CoMAND method (Conformational Mapping by Analytical NOESY Decomposition<sup>27</sup>). This method analyzes a 3D CNH-NOESY spectrum (i.e. 3D  $[H]C,NH$  HSQC-NOESY-HSQC), which displays only NOE contacts between  $^{15}N$ -bound and  $^{13}C$ -bound protons and is therefore

intrinsically diagonal free<sup>65</sup>. As a first step, one-dimensional <sup>13</sup>C sub-spectra (strips) were extracted from this spectrum ( $\tau_m=400$  ms). Each strip is taken perpendicular to a specific <sup>15</sup>N-HSQC position and represents contacts to a single <sup>15</sup>N-bound proton, edited by the <sup>13</sup>C shift of the attached carbon. For (P3-7)<sub>2</sub> we obtained strips for 18 main chain amide protons (residues L7 to K24) and all 4 glutamine side chain H<sub>ε21</sub> protons (Q11, Q14, Q17, Q20). These strips are analyzed in terms of a quantitative R-factor expressing the agreement between experimental and back-calculated spectra. Global back-calculation parameters for CoMAND were optimized by grid searching, resulting in an overall correlation time of 2.0 ns and effective <sup>13</sup>C signal halfwidth of 14 Hz.

For a reconstruction of the experimental <sup>13</sup>C strips, CoMAND compiles a linear combination of strips back-calculated from a set of trial conformers that should reflect the conformational space of each residue. Here, we used the a99sb-disp and DES-amber MD trajectories and back-calculated 13002 frames from each trajectory for each of the 22 experimental strips. For each residue, the conformational ensemble producing the lowest R-factor was then compiled using the CoMAND stochastic optimisation method<sup>27</sup>. A starting conformer is randomly selected from the 20 conformers with lowest R-factors. All conformers are then tested in random order, with a new member added to the ensemble if it decreases the R-factor by more than a given threshold (0.0005). Convergence is achieved if no further conformer is found or if the ensemble reaches a maximum size, here set at 20 structures. Due to its stochastic nature, this selection procedure can be repeated to produce ensembles with similar R-factors, but sampling a wider range of conformers.

For (P3-7)<sub>2</sub> we applied a two-step protocol for each residue. In the first step we established the minimum R-factor by compiling 100 per-residue ensembles, optimizing over single experimental strips. These per-residue ensembles were also used to define a set of *witness* strips; i.e. those whose R-factors may be affected by conformational changes in the residue in question. In the second step, we obtained the conformational distribution by co-optimizing over these witness strips. For each set, 100 optimization trials were run for each MD trajectory frame pool, resulting in 100 ensembles per frame pool, each typically containing 5 to 15 members. After removing ensembles with R-factors significantly above average (90% confidence interval), the set of conformers used in co-optimization (typically over 1000) was pooled to represent the conformational diversity for each residue.

To quantify the conformational distributions, we clustered the data via Gaussian mixture modeling (GMM). A vector of  $n$  features - here dihedral angles - was defined for each conformer which was then used to train a model describing the probability  $p(k|x)$  that a data

point  $x$  is a member of cluster  $k$ . For each cluster, this probability is defined by an  $n$ -dimensional multivariate Gaussian distribution representing its center and shape, and by a prior probability,  $p(k)$ , corresponding to its relative population. These model parameters were fitted to the training data using the Expectation Maximization (EM) algorithm, modified to accommodate the periodic nature of dihedral angles. For  $(P3-7)_2$ , we applied the GMM method for the  $\chi_1/\chi_2$  pairs of all leucine and glutamine residues. As the number of clusters that best describe the training data was not known *a priori*, we searched values from 1 to 9 systematically and assessed the fit via the Bayesian information criterion, a measure that includes a penalty for model complexity. EM initialisation requires an arbitrary seed value for each cluster center. For  $\chi_1/\chi_2$  pairs, it is convenient to select seeds at the center of a rotameric form, with seeds progressively added with increasing cluster number, according to their database frequency. The best scoring model was stored for each residue.

The GMM method provides a compact but detailed description of conformational landscapes for use in downstream calculations. Here we have applied it to Monte-Carlo conformational sampling as part of an extended “greedy” R-factor optimisation protocol. The pooled a99sb-disp and DES-amber MD trajectories were systematically sampled and the conformer affording the greatest reduction in global R-factor added in each iteration. Thus, 2-4 conformers were typically added to the ensemble, which was then further optimized by adjusting the side chain conformations for leucine and glutamine residues using  $\chi_1/\chi_2$  combinations from the corresponding GMM model with a 0.05 probability cutoff. For each residue in the ensemble, up to 30  $\chi_1/\chi_2$  combinations were tested to find sterically acceptable conformations lowering the global R-factor. For glutamine, the  $\chi_3$  angle was additionally sampled around population centers pertaining to each  $\chi_1/\chi_2$  combination (5 trials; standard deviation  $8^\circ$ ). Note that enthalpic contributions from hydrogen bonding were not considered in testing conformers, such that their selection was primarily driven by the reduction in R-factors. This iterative process of ensemble selection and modification was repeated until no further conformers were added by the greedy step. This protocol was repeated 20 times to probe the consistency of results and an example chosen as the final ensemble (Fig. 2g).

### *Molecular Dynamics simulations*

The trajectories for the  $(P3-7)_2$  peptide used in the CoMAND analysis and for the TBP peptide shown in Fig. 5h were generated with the Well Tempered Ensemble (WTE)<sup>39,70</sup> enhanced sampling algorithm. We used 26 energy biased replicas within the temperature range from 275 K to 500 K, and two unbiased replicas at 278 K and 298 K. The unbiased replica at 278 K (at which temperature the 3D  $[H]C,NH$  HSQC-NOESY-HSQC was

measured) was then used to generate the conformations for our CoMAND analysis and Fig 5h. For the biased replicas, the energy bias was increased during the first 500 ps and then kept constant. All replicas were subsequently used in a production simulation for 200 ns, where conformations used by CoMAND were extracted from the last 150 ns. Exchange between replicas was monitored to ensure good replica diffusion in temperature space. The production simulation was run in the NPT ensemble. We used two recent force fields for these simulations, DES-amber<sup>28</sup> and a99sb-disp<sup>29</sup>, each with its pertaining TIP4P-D water model. Bonds with hydrogen atoms were constrained and a time-step of 2 fs was used. For our simulations we used Gromacs 2019.4<sup>71–73</sup> patched with the PLUMED library<sup>74</sup> version 2.5.3<sup>75</sup> to enable the WTE sampling method.

Two sets of trajectories for segments of L<sub>3</sub>X peptides (see Table S1) with the sequence L<sup>1</sup>L<sup>2</sup>L<sup>3</sup>XQ<sup>1</sup>Q<sup>2</sup>Q<sup>3</sup>Q<sup>4</sup>Q<sup>5</sup>Q<sup>6</sup>Q<sup>7</sup>Q<sup>8</sup> (where X is any of the 20 natural proteinogenic amino acids) were calculated using either the Charmm36m force field<sup>33</sup> with a TIP3P<sup>76</sup> water model or the a99sb-disp force field<sup>29</sup> with its TIP4P-D water model. To bias the simulations towards relevant conformations, calculations were started from a fully helical conformation and the backbone  $\phi$  and  $\psi$  dihedral angles of residues L<sup>2</sup> to Q<sup>5</sup> were restrained to -60° and -40°, respectively, with a spring constant  $k$  value of 5 kJ mol<sup>-1</sup> degree<sup>-2</sup>. Similarly, a bias to increase the occurrence of Q<sup>4</sup> → X side chain to main chain hydrogen bonds was introduced by restraining the distance between the main chain O of residue X and the N<sub>ε2</sub> of residue Q<sup>4</sup> to 4 Å, with a spring constant  $k$  value of 50 kJ mol<sup>-1</sup> nm<sup>-2</sup>. Structure minimization and thermalization at 278 K was performed in the NVT ensemble for 1 ns. For the 1 μs NPT production runs we used Gromacs 2019.4<sup>71–73</sup> and an equilibration period of 100 ns that was excluded from trajectory analysis.

A trajectory was calculated for peptide Q1E using the CHARMM22\*<sup>77</sup> force field and TIP3P water model<sup>76</sup> to obtain different structures with i-4 → i side chain to main chain hydrogen bonds involving the glutamic acid carboxyl group that served as seeds for the QM/MM simulations (see below). The dihedral angle  $\chi_4$  orienting the glutamic acid carboxyl group was restrained to 0° since its most stable conformation in CHARMM22\*, corresponding to  $\chi_4 = 180^\circ$ , is incompatible with the experimentally indicated side chain to main chain interaction. A fully helical starting structure was thermalized (300K) and equilibrated in the NVT ensemble for 1 ns. The 1 μs production run was obtained using ACEMD<sup>78</sup> with a 100 ns equilibration period.

## *QM/MM simulations*

Four starting structures were selected from the classical MD trajectory of peptide Q1E conserving their box of water and ions. For the QM/MM simulations we used AMBER 20<sup>79</sup> coupled to the QM Terachem 1.9 interface<sup>80–82</sup>. The QM subsystem was described at the BLYP/6-31G\* level of theory including dispersion corrections<sup>83</sup> and comprised 66 atoms including linker atoms. The classical subsystem was treated with the CHARMM22\*<sup>77</sup> and TIP3P<sup>76</sup> force fields. The linker atom procedure was employed to saturate the valency of the frontier atoms and electrostatic embedding was used as implemented in AMBER. An electrostatic cutoff of 12 Å and periodic boundary conditions were employed throughout all QM/MM MD simulations, using a time step of 1 fs. Structures were minimized, thermalized, and equilibrated for 10 ps at the QM/MM level prior to the 150 ps-long production runs. Finally, for each of the 150 ps QM/MM-MD runs, a Natural Bond Critical Point analysis<sup>84,85</sup> was performed using NBO 7.0<sup>86</sup>.

## *Database motif searches*

UniprotKB<sup>34</sup> (including both the Swiss-Prot and the TrEMBL databases) was queried for protein sequences containing motifs that fulfill our design rules using the ScanProsite<sup>35</sup> motif search tool hosted at the Expasy website (<https://prosite.expasy.org/scanprosite/>). The query motif was introduced in Prosite format as:

[LFYWIM]-X-X-[LFYWIM]-Q-X-[LFYWIM]-Q-X-[LFYWIM]-Q-X-Q-X-X-Q-X-X to quest for sequences with  $n = 4 Q_{i+4} - [LFYWIM]_i$  pairs, with the number of central -[LFYWIM]-Q-X- triplets increased stepwise for concomitant increases in the requested number of  $Q_{i+4} - [LFYWIM]_i$  pairs. UniprotKB annotation scores and taxonomic lineage information was obtained by programmatically accessing this information in UniprotKB using the accession codes from the ScanProsite searches. The same codes were used to retrieve structural models from the AlphaFold Database<sup>18</sup>, in which the secondary structure of the (P3-7)<sub>n</sub>-like motifs fulfilling our design rules was obtained using DSSP<sup>36</sup>.

## *Hydrogen bond criteria*

We considered two atoms to be hydrogen bonded if the distance between donor H and the acceptor O was shorter than 2.4 Å and their angle was larger than 120°.

## **Acknowledgements**

We thank Borja Mateos, Ben Lehner, Ernest Giralt and Luis Serrano for helpful discussions and the ICTS NMR facility, managed by the scientific and technological centers of the

University of Barcelona (CCiT UB), for their help in NMR. X.S. acknowledges funding from AGAUR (2017 SGR 324), MINECO (BIO2015-70092-R and PID2019-110198RB-I00), and the European Research Council (CONCERT, contract number 648201). B.B.K acknowledges funding from the Novo Nordisk Foundation (#NNF18OC0033926). MO acknowledges funding from the Instituto Nacional de Bioinformática, The EU BioExcel Centre of Excellence for HPC and the Spanish Ministry of Science (RTI2018-096704-B-100) and the Instituto de Salud Carlos III–Instituto Nacional de Bioinformática (ISCIII PT 17/0009/0007 co-funded by the Fondo Europeo de Desarrollo Regional). MO is an ICREA Academy scholar and JA is a Juan de la Cierva fellow. IRB Barcelona is the recipient of a Severo Ochoa Award of Excellence from MINECO (Government of Spain).

## References

1. Jones, S. & Thornton, J. M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 13–20 (1996).
2. Fuller, J. C., Burgoyne, N. J. & Jackson, R. M. Predicting druggable binding sites at the protein–protein interface. *Drug Discov. Today* **14**, 155–161 (2009).
3. Tietz, O., Cortezon-Tamarit, F., Chalk, R., Able, S. & Vallis, K. A. Tricyclic cell-penetrating peptides for efficient delivery of functional antibodies into cancer cells. *Nat. Chem.* **14**, 284–293 (2022).
4. Philippe, G. J. B., Craik, D. J. & Henriques, S. T. Converting peptides into drugs targeting intracellular protein-protein interactions. *Drug Discov. Today* **26**, 1521–1531 (2021).
5. Ran, X. & Gestwicki, J. E. Inhibitors of protein–protein interactions (PPIs): an analysis of scaffold choices and buried surface area. *Curr. Opin. Chem. Biol.* **44**, 75–86 (2018).
6. Rezaei Araghi, R. & Keating, A. E. Designing helical peptide inhibitors of protein-protein interactions. *Curr. Opin. Struct. Biol.* **39**, 27–38 (2016).
7. Cromm, P. M., Spiegel, J. & Grossmann, T. N. Hydrocarbon stapled peptides as modulators of biological function. *ACS Chem. Biol.* **10**, 1362–1375 (2015).
8. Pelay-Gimeno, M., Glas, A., Koch, O. & Grossmann, T. N. Structure-Based Design of Inhibitors of Protein-Protein Interactions: Mimicking Peptide Binding Epitopes. *Angew.*

- Chem. Int. Ed Engl.* **54**, 8896–8927 (2015).
9. Hilinski, G. J. *et al.* Stitched  $\alpha$ -helical peptides via bis ring-closing metathesis. *J. Am. Chem. Soc.* **136**, 12314–12322 (2014).
  10. Chang, Y. S. *et al.* Stapled  $\alpha$ -helical peptide drug development: a potent dual inhibitor of MDM2 and MDMX for p53-dependent cancer therapy. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E3445–54 (2013).
  11. Escobedo, A. *et al.* Side chain to main chain hydrogen bonds stabilize a polyglutamine helix in a transcription factor. *Nat. Commun.* **10**, 2034 (2019).
  12. Eftekharzadeh, B. *et al.* Sequence Context Influences the Structure and Aggregation Behavior of a PolyQ Tract. *Biophys. J.* **110**, 2361–2366 (2016).
  13. Urbanek, A. *et al.* Flanking Regions Determine the Structure of the Poly-Glutamine in Huntingtin through Mechanisms Common among Glutamine-Rich Human Proteins. *Structure* **28**, 733–746 (2020).
  14. Nielsen, J. T. & Mulder, F. A. A. CheSPI: chemical shift secondary structure population inference. *J. Biomol. NMR* **75**, 273–291 (2021).
  15. Nielsen, J. T. & Mulder, F. A. A. POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins. *J. Biomol. NMR* **70**, 141–165 (2018).
  16. Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
  17. Reid Alderson, T., Pritišanac, I., Moses, A. M. & Forman-Kay, J. D. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *bioRxiv* 2022.02.18.481080 (2022) doi:10.1101/2022.02.18.481080.
  18. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
  19. Baias, M. *et al.* Structure and Dynamics of the Huntingtin Exon-1 N-Terminus: A Solution NMR Perspective. *J. Am. Chem. Soc.* **139**, 1168–1176 (2017).

20. Hong, J.-Y. *et al.* Structural and dynamic studies reveal that the Ala-rich region of ataxin-7 initiates  $\alpha$ -helix formation of the polyQ tract but suppresses its aggregation. *Sci. Rep.* **9**, 7481 (2019).
21. Kosol, S. *et al.* Interaction between the scaffold proteins CBP by IQGAP1 provides an interface between gene expression and cytoskeletal activity. *Sci. Rep.* **10**, 1–15 (2020).
22. Sicorello, A., Różycki, B., Konarev, P. V., Svergun, D. I. & Pastore, A. Capturing the Conformational Ensemble of the Mixed Folded Polyglutamine Protein Ataxin-3. *Structure* **29**, 70–81.e5 (2021).
23. Muñoz, V. & Serrano, L. Elucidating the folding problem of helical peptides using empirical parameters. *Nat. Struct. Biol.* **1**, 399–409 (1994).
24. Sheu, S.-Y., Yang, D.-Y., Selzle, H. L. & Schlag, E. W. Energetics of hydrogen bonds in peptides. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12683–12687 (2003).
25. Neal, S., Nip, A. M., Zhang, H. & Wishart, D. S. Rapid and accurate calculation of protein <sup>1</sup>H, <sup>13</sup>C and <sup>15</sup>N chemical shifts. *J. Biomol. NMR* **26**, 215–240 (2003).
26. Shapovalov, M. V. & Dunbrack, R. L., Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**, 844–858 (2011).
27. ElGamacy, M., Riss, M., Zhu, H., Truffault, V. & Coles, M. Mapping Local Conformational Landscapes of Proteins in Solution. *Structure* **27**, 853–865.e5 (2019).
28. Piana, S., Robustelli, P., Tan, D., Chen, S. & Shaw, D. E. Development of a Force Field for the Simulation of Single-Chain Proteins and Protein–Protein Complexes. *J. Chem. Theory Comput.* **16**, 2494–2507 (2020).
29. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4758–E4766 (2018).
30. Pace, C. N. & Scholtz, J. M. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* **75**, 422–427 (1998).

31. Wan, W. Y. & Milner-White, E. J. A recurring two-hydrogen-bond motif incorporating a serine or threonine residue is found both at alpha-helical N termini and in other situations. *J. Mol. Biol.* **286**, 1651–1662 (1999).
32. Luque, I., Mayorga, O. L. & Freire, E. Structure-based thermodynamic scale of alpha-helix propensities in amino acids. *Biochemistry* **35**, 13681–13688 (1996).
33. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
34. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
35. de Castro, E. *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* **34**, W362–5 (2006).
36. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
37. Baker, E. G. *et al.* Local and macroscopic electrostatic interactions in single  $\alpha$ -helices. *Nat. Chem. Biol.* **11**, 221–228 (2015).
38. Bravo-Arredondo, J. M. *et al.* The folding equilibrium of huntingtin exon 1 monomer depends on its polyglutamine tract. *J. Biol. Chem.* **293**, 19613–19623 (2018).
39. Bonomi, M. & Parrinello, M. Enhanced sampling in the well-tempered ensemble. *Phys. Rev. Lett.* **104**, 190601 (2010).
40. De Mol, E. *et al.* Regulation of Androgen Receptor Activity by Transient Interactions of Its Transactivation Domain with General Transcription Regulators. *Structure* **26**, 145–152.e3 (2018).
41. Kamada, K., Roeder, R. G. & Burley, S. K. Molecular mechanism of recruitment of TFIIIF- associating RNA polymerase C-terminal domain phosphatase (FCP1) by transcription factor IIF. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 2296–2299 (2003).
42. Yang, A. *et al.* NMR Structure of a Complex Formed by the Carboxyl-Terminal Domain

- of Human RAP74 and a Phosphorylated Peptide from the Central Domain of the FCP1 Phosphatase †‡. *Biochemistry* **48**, 1964–1974 (2009).
43. Simm, D. & Kollmar, M. Waggawagga-CLI: A command-line tool for predicting stable single  $\alpha$ -helices (SAH-domains), and the SAH-domain distribution across eukaryotes. *PLoS One* **13**, e0191924 (2018).
  44. Swanson, C. J. & Sivaramakrishnan, S. Harnessing the unique structural properties of isolated  $\alpha$ -helices. *J. Biol. Chem.* **289**, 25460–25467 (2014).
  45. Süveges, D., Gáspári, Z., Tóth, G. & Nyitrai, L. Charged single alpha-helix: a versatile protein structural motif. *Proteins* **74**, 905–916 (2009).
  46. Song, Z. *et al.* Modulation of polypeptide conformation through donor-acceptor transformation of side-chain hydrogen bonding ligands. *Nat. Commun.* **8**, 92 (2017).
  47. Harrington, L., Fletcher, J. M., Heermann, T., Woolfson, D. N. & Schwille, P. De novo design of a reversible phosphorylation-dependent switch for membrane targeting. *Nat. Commun.* **12**, 1472 (2021).
  48. Sia, S. K. & Kim, P. S. Protein grafting of an HIV-1-inhibiting epitope. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9756–9761 (2003).
  49. Au, Y. Z., Wang, T., Sigua, L. H. & Qi, J. Peptide-Based PROTAC: The Predator of Pathological Proteins. *Cell chemical biology* vol. 27 637–639 (2020).
  50. Faure, A. J. *et al.* Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).
  51. Orr, H. T. & Zoghbi, H. Y. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.* **30**, 575–621 (2007).
  52. Warner, J. B., 4th *et al.* Monomeric Huntingtin Exon 1 Has Similar Overall Structural Features for Wild-Type and Pathological Polyglutamine Lengths. *J. Am. Chem. Soc.* **139**, 14456–14469 (2017).
  53. Urbanek, A. *et al.* Evidence of the reduced abundance of proline cis conformation in protein poly proline tracts. *J. Am. Chem. Soc.* **142**, 7976–7986 (2020).

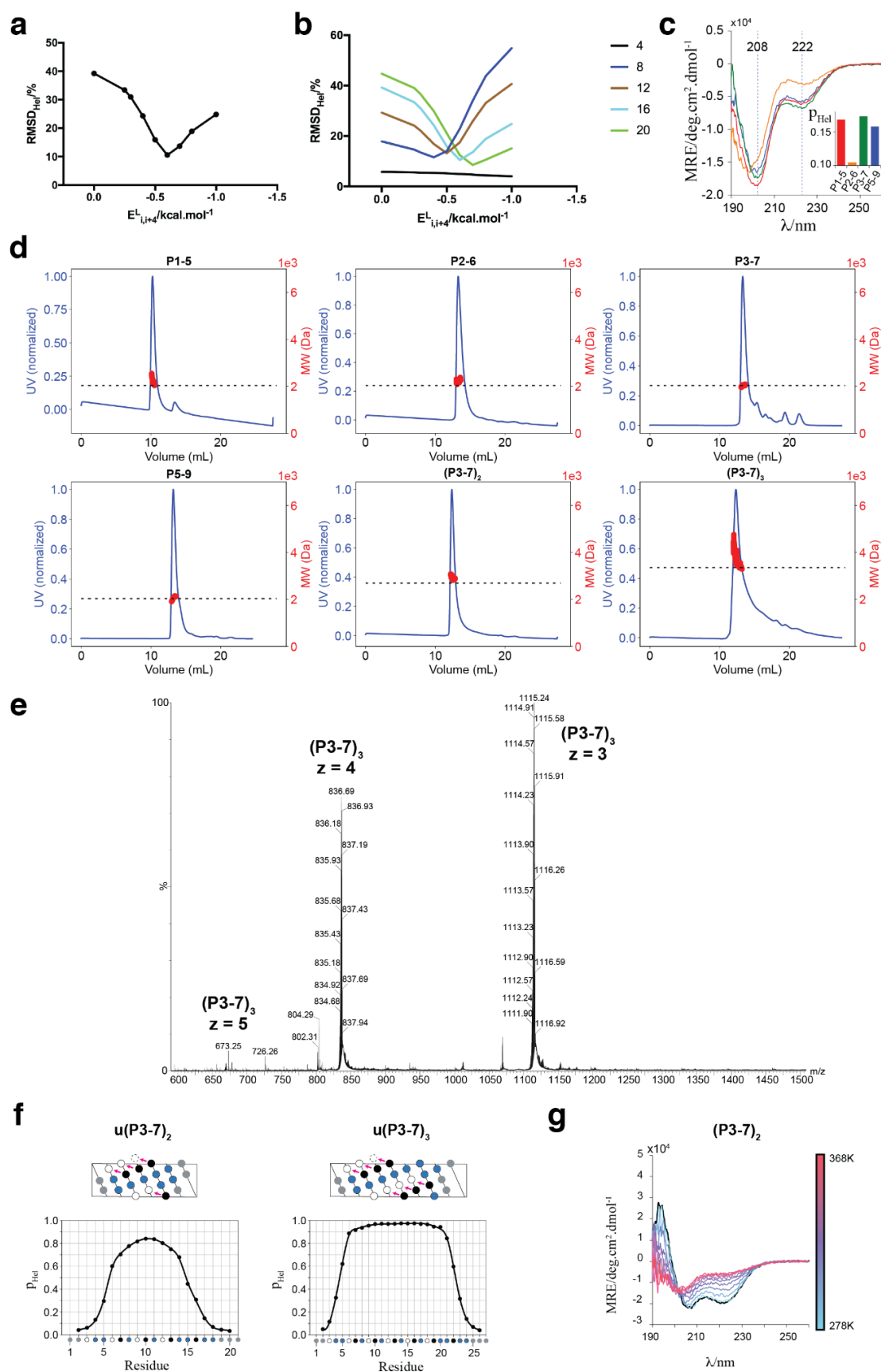
54. Urbanek, A. *et al.* A General Strategy to Access Structural Information at Atomic Resolution in Polyglutamine Homorepeats. *Angew. Chem. Int. Ed Engl.* **130**, 3660–3663 (2018).
55. Kotler, S. A. *et al.* Probing initial transient oligomerization events facilitating Huntingtin fibril nucleation at atomic resolution by relaxation-based NMR. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 3562–3571 (2019).
56. Jayaraman, M. *et al.* Slow amyloid nucleation via  $\alpha$ -helix-rich oligomeric intermediates in short polyglutamine-containing huntingtin fragments. *J. Mol. Biol.* **415**, 881–899 (2012).
57. Escobedo, A., Chiesa, G. & Salvatella, X. Recombinant Production of Monomeric Isotope-Enriched Aggregation-Prone Peptides: Polyglutamine Tracts and Beyond. *Methods Mol. Biol.* **2141**, 211–231 (2020).
58. Hwang, T. L. & Shaka, A. J. Water Suppression That Works. Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *J. Magn. Reson. A* **112**, 275–279 (1995).
59. Solyom, Z. *et al.* BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. *J. Biomol. NMR* **55**, 311–321 (2013).
60. Bermel, W. *et al.* Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins. *J. Am. Chem. Soc.* **128**, 3918–3919 (2006).
61. Orekhov, V. Y. & Jaravine, V. A. Analysis of non-uniformly sampled spectra with multi-dimensional decomposition. *Prog. Nucl. Magn. Reson. Spectrosc.* **59**, 271–292 (2011).
62. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
63. Vranken, W. F. *et al.* The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* **59**, 687–696 (2005).
64. Platzer, G., Okon, M. & McIntosh, L. P. pH-dependent random coil (1)H, (13)C, and (15)N chemical shifts of the ionizable amino acids: a guide for protein pK<sub>a</sub>

- measurements. *J. Biomol. NMR* **60**, 109–129 (2014).
65. Diercks, T., Coles, M. & Kessler, H. An efficient strategy for assignment of cross-peaks in 3D heteronuclear NOESY experiments. *J. Biomol. NMR* **15**, 177–180 (1999).
  66. Liu, A. *et al.* Hydrogen-bond detection, configuration assignment and rotamer correction of side-chain amides in large proteins by NMR spectroscopy through protium/deuterium isotope effects. *Chembiochem* **9**, 2860–2871 (2008).
  67. Ferrage, F., Reichel, A., Battacharya, S., Cowburn, D. & Ghose, R. On the measurement of  $^{15}\text{N}\{-^1\text{H}\}$  nuclear Overhauser effects. 2. Effects of the saturation scheme and water signal suppression. *J. Magn. Reson.* **207**, 294–303 (2010).
  68. Chou, J. J., Gaemers, S., Howder, B., Louis, J. M. & Bax, A. A simple apparatus for generating stretched polyacrylamide gels, yielding uniform alignment of proteins and detergent micelles. *J. Biomol. NMR* **21**, 377–382 (2001).
  69. Zweckstetter, M. NMR: prediction of molecular alignment from structure using the PALES software. *Nat. Protoc.* **3**, 679–690 (2008).
  70. Deighan, M., Bonomi, M. & Pfaendtner, J. Efficient Simulation of Explicitly Solvated Proteins in the Well-Tempered Ensemble. *J. Chem. Theory Comput.* **8**, 2189–2192 (2012).
  71. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56 (1995).
  72. Van Der Spoel, D. *et al.* GROMACS: fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718 (2005).
  73. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25 (2015).
  74. PLUMED consortium. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **16**, 670–673 (2019).
  75. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New

- feathers for an old bird. *Comput. Phys. Commun.* **185**, 604–613 (2014).
76. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* vol. 79 926–935 (1983).
  77. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **100**, L47–9 (2011).
  78. Harvey, M. J., Giupponi, G. & Fabritiis, G. D. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.* **5**, 1632–1639 (2009).
  79. Case, D. A. *et al.* *AMBER 2020, University of California, San Francisco.* (2020).
  80. Ufimtsev, I. S. & Martínez, T. J. Quantum Chemistry on Graphical Processing Units. 1. Strategies for Two-Electron Integral Evaluation. *J. Chem. Theory Comput.* **4**, 222–231 (2008).
  81. Ufimtsev, I. S. & Martinez, T. J. Quantum Chemistry on Graphical Processing Units. 2. Direct Self-Consistent-Field (SCF) Implementation. *Journal of Chemical Theory and Computation* vol. 5 3138–3138 (2009).
  82. Ufimtsev, I. S. & Martinez, T. J. Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *J. Chem. Theory Comput.* **5**, 2619–2628 (2009).
  83. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).
  84. Weinhold, F. Natural Bond Critical Point analysis: quantitative relationships between natural bond orbital-based and QTAIM-based topological descriptors of chemical bonding. *J. Comput. Chem.* **33**, 2440–2449 (2012).
  85. Bader, R. F. W. Atoms in molecules. *Accounts of Chemical Research* vol. 18 9–15 (1985).
  86. Glendening, E. D. *et al.* *NBO 7.0.* (2018).

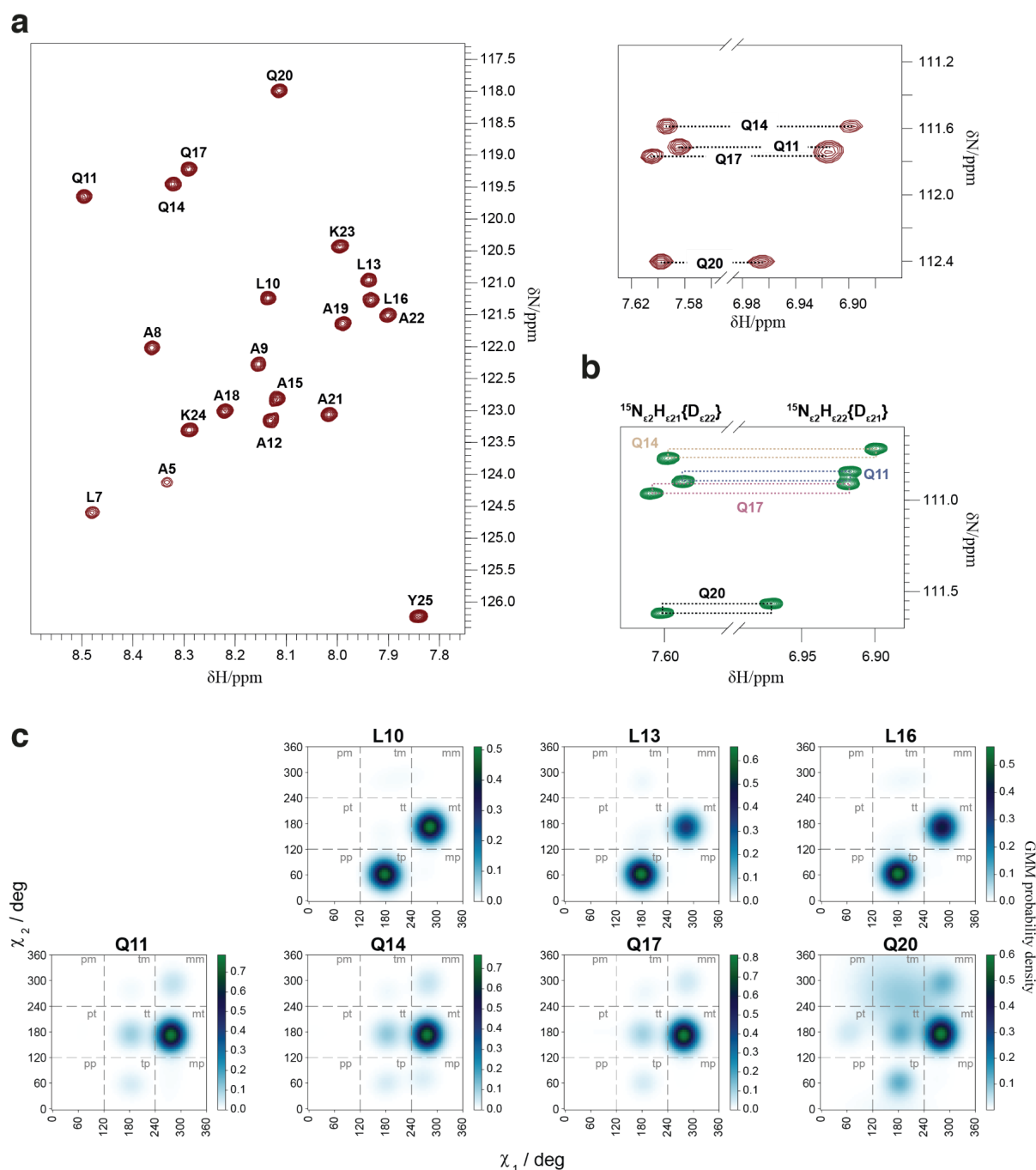


## Supplementary Figures

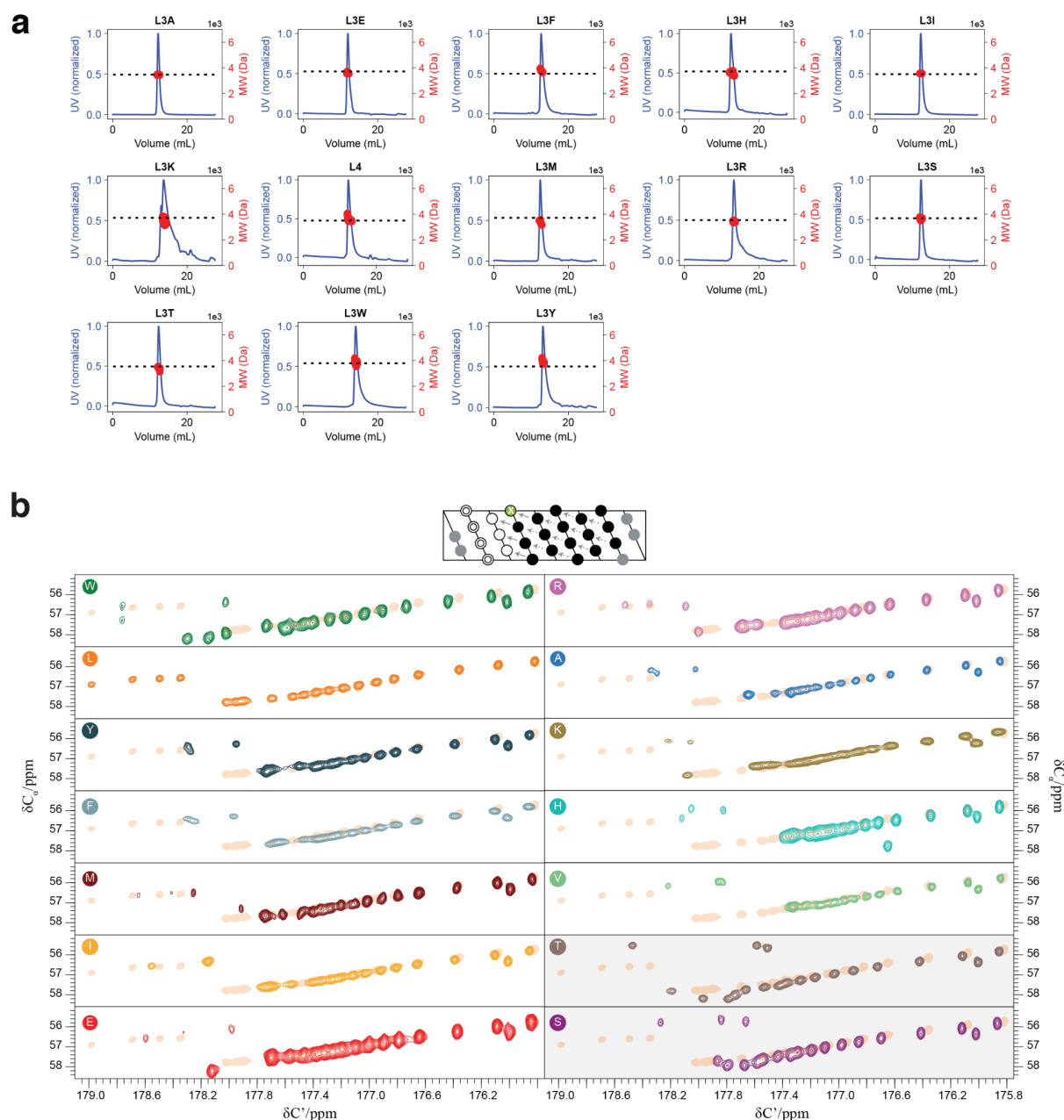


**Fig. S1 | Cooperativity of  $\text{Gln}_{i+4} \rightarrow \text{Leu}_i$  bifurcated hydrogen bonds in monomeric polyQ tracts and designed sequences.**

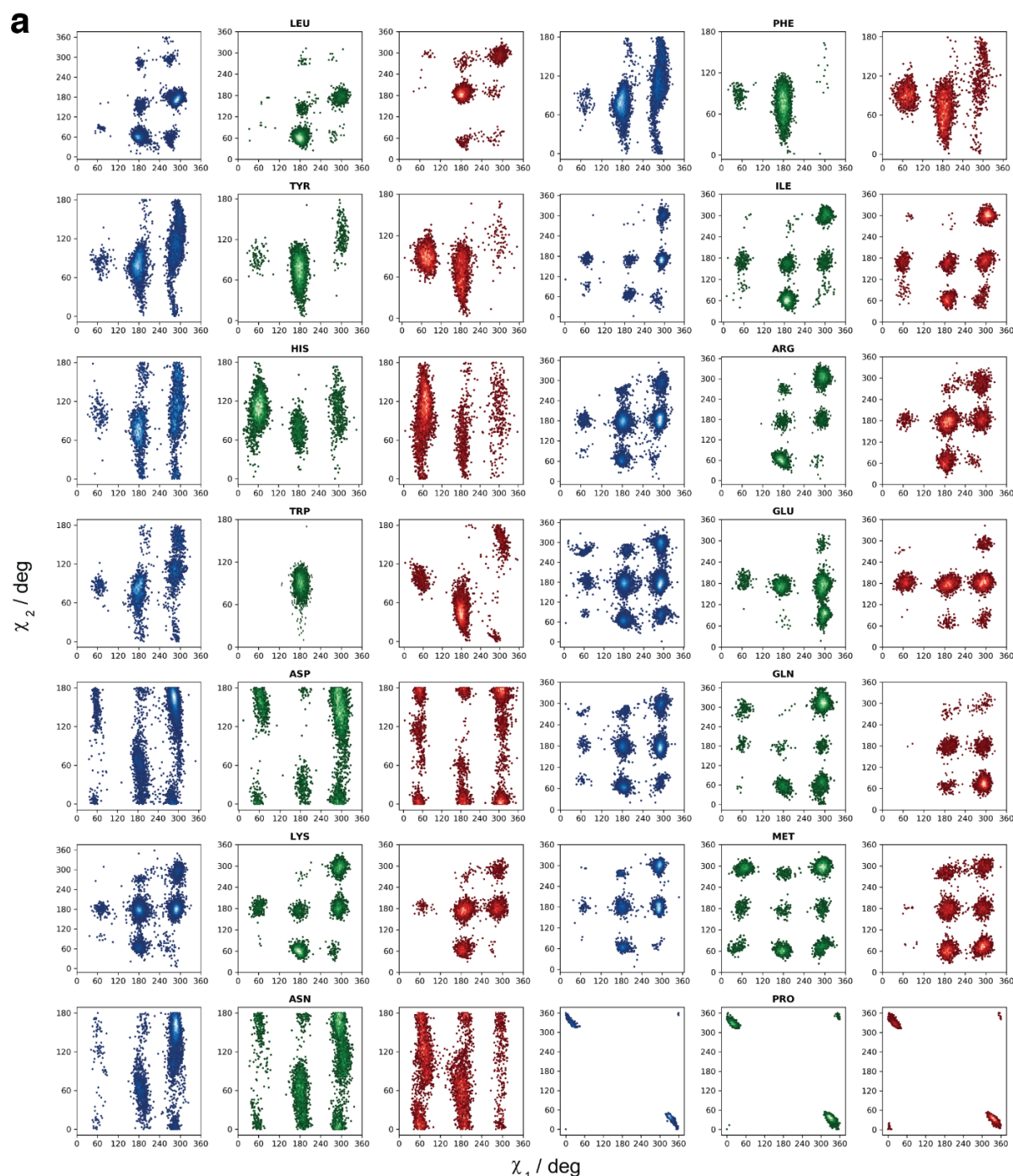
**a** Optimizing the value of a newly introduced energy term  $E_{i,i+4}^L$  in *Agadir* to account for the  $\text{Gln}_{i+4} \rightarrow \text{Leu}_i$  interaction improves its accuracy at predicting the helicity of  $\text{L}_4\text{Q}_{16}$  (Table S1). **b** The optimal value of  $E_{i,i+4}^L$  depends on polyQ tract length. Numbers in the legend indicate the length of the  $\text{L}_4\text{Q}_n$  peptides (full sequence  $\text{KKPGASL}_4\text{Q}_n\text{KKY}$ ). **c** Overlay of the CD spectra obtained for the peptides in Fig. 1a. The dashed vertical lines highlight the positions where minima are observed in the CD spectrum of the  $\alpha$ -helix (208 and 222 nm). In the inset, average peptide helicities are shown as obtained from deconvoluting the CD curve using the algorithm CONTIN (reference set 7) hosted at DichroWeb (<http://dichroweb.cryst.bbk.ac.uk>). **d** SEC-MALS analyses of the peptides in Fig. 1a and 1b. The horizontal dashed line indicates the molecular weight of the monomeric species calculated using the ProtParam algorithm hosted at Expasy (<https://web.expasy.org/protparam/>). **e** Native mass spectrum of a  $^{13}\text{C}$ - $^{15}\text{N}$  labeled sample of the peptide  $(\text{P3-7})_3$ , as obtained by Q-TOF MS using a Synapt G1-HDMS (see supplementary methods). Three ionization states corresponding to the monomeric species were independently detected. No oligomeric species were detected. **f** Chemical shifts-derived helical propensity ( $p_{\text{Hel}}$ ) profiles of the  $\text{u}(\text{P3-7})_2$  and  $\text{u}(\text{P3-7})_3$  peptides, which are versions of the peptides shown in Fig. 1b lacking the PGAS N-capping sequence (the 'u' stands for *uncapped*). **g** Thermal unfolding of  $(\text{P3-7})_2$ . Shown are CD spectra acquired every 10 K in the 278 K - 368 K range (blue to red color gradient). The CD spectrum back at 278 K after thermal unfolding is shown in black.



**Fig. S2 | Structural features of Gln<sub>i+4</sub>→Leu<sub>i</sub> bifurcated hydrogen bonds: supporting analyses of the peptide (P3-7)<sub>2</sub>.** **a** Assigned <sup>1</sup>H-<sup>15</sup>N HSQC spectrum of the peptide (P3-7)<sub>2</sub>. Left: region of the spectrum displaying main chain N-H correlations. The spectrum shows wide signal dispersion in both dimensions and no signal overlap, which was exploited for the structural characterization shown in Fig. 2. Right: region of the spectrum showing the Q side chain NH<sub>2</sub> correlations. **b** Side chain region of a <sup>1</sup>H-<sup>15</sup>N HSQC spectrum acquired on a (P3-7)<sub>2</sub> sample in 50% D<sub>2</sub>O applying deuterium decoupling during the <sup>15</sup>N evolution time. The differential deuterium isotope shift in the NHD species results in different <sup>15</sup>N chemical shifts for each of them, which allows the unambiguous stereospecific assignment of the side chain carboxamide H<sub>ε</sub> resonances<sup>66</sup> **c** Probability densities for χ<sub>1</sub> and χ<sub>2</sub> of all residues captured by the Gaussian Mixture Model (GMM) from the CoMAND per-residue fits.

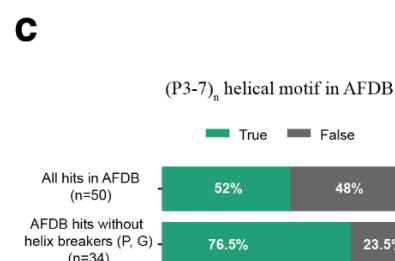
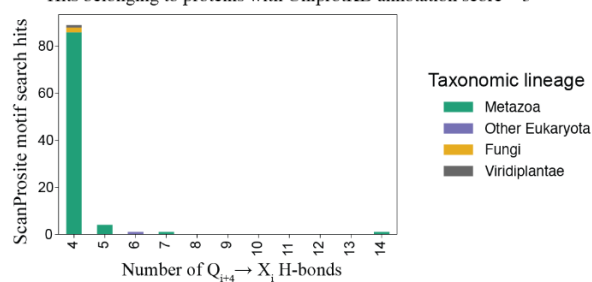


**Fig. S3 | SEC-MALS analyses and CACO spectra of the L<sub>3</sub>XQ<sub>16</sub> peptides. a** SEC-MALS analyses of the peptides in the L<sub>3</sub>XQ<sub>16</sub> series shown in Fig. 3. The horizontal dashed line indicates the molecular weight of the monomeric species calculated using the ProtParam algorithm hosted at ExPASy (<https://web.expasy.org/protparam/>). **b** <sup>13</sup>C-detected CACO spectra of the L<sub>3</sub>XQ<sub>16</sub> variants, sorted by their observed helicity, with the spectrum of L<sub>4</sub>Q<sub>16</sub> underlaid (orange shadow) in every panel for reference. Spectra shaded in gray indicate variants with outlier behavior.

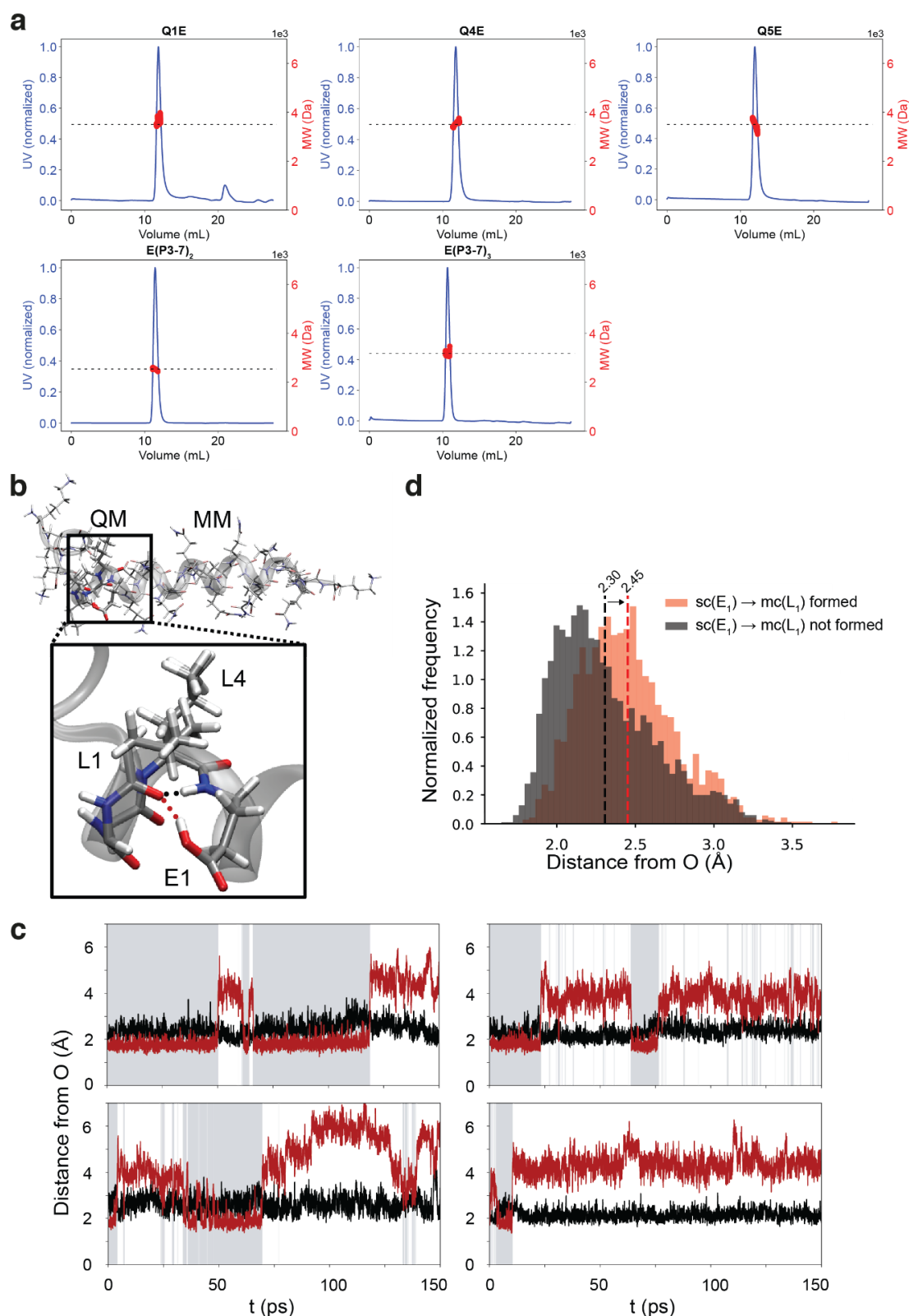


**b**

$\Omega$ -X-X-( $\Omega$ -Q-X)<sub>n-1</sub>-X-Q-X-X motif in Swiss-Prot and TrEMBL databases  
 Hits belonging to proteins with UniprotKB annotation score > 3

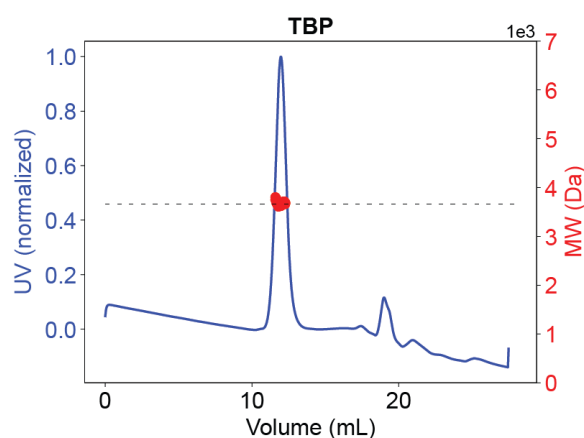


**Fig. S4 | Rotamer sampling in the L<sub>3</sub>XQ<sub>16</sub> MD simulations and supporting data on the natural sequences fulfilling our design rules.** **a** Side chain  $\chi_1$ - $\chi_2$  dihedral angle correlations for residues with at least three side chain carbon atoms. In blue, values as reported in the BBRep database for residues located in the context of a helix. In green, values sampled throughout the Charmm36m 1  $\mu$ s MD trajectories calculated for the corresponding L<sub>3</sub>XQ<sub>16</sub> variants. In red, values sampled throughout the a99sb-disp 1  $\mu$ s MD trajectories calculated for the corresponding L<sub>3</sub>XQ<sub>16</sub> variants. **b** Number of ScanProsite-identified<sup>35</sup> protein sequences with an UniprotKB<sup>34</sup> annotation score > 3 found in the Swiss-Prot and TrEMBL databases containing the (P3-7)<sub>n</sub> motif ( $\Omega$  = W, L, Y, F, I and M; X = any amino acid). Counts are grouped by taxonomic lineage as obtained from UniProt. **c** 50 proteins containing sequence motifs fulfilling our design rules have models in the AlphaFold Database (AFDB)<sup>18</sup>. The top bar shows the percentage of these proteins where the (P3-7)<sub>n</sub>-like motif is predicted to be helical from residue 1 to residue -4, as obtained from a DSSP analysis<sup>36</sup> of the structural models. The bottom bar shows the same for proteins not containing helix breaking residues (P, G) in the same segment (n=34).

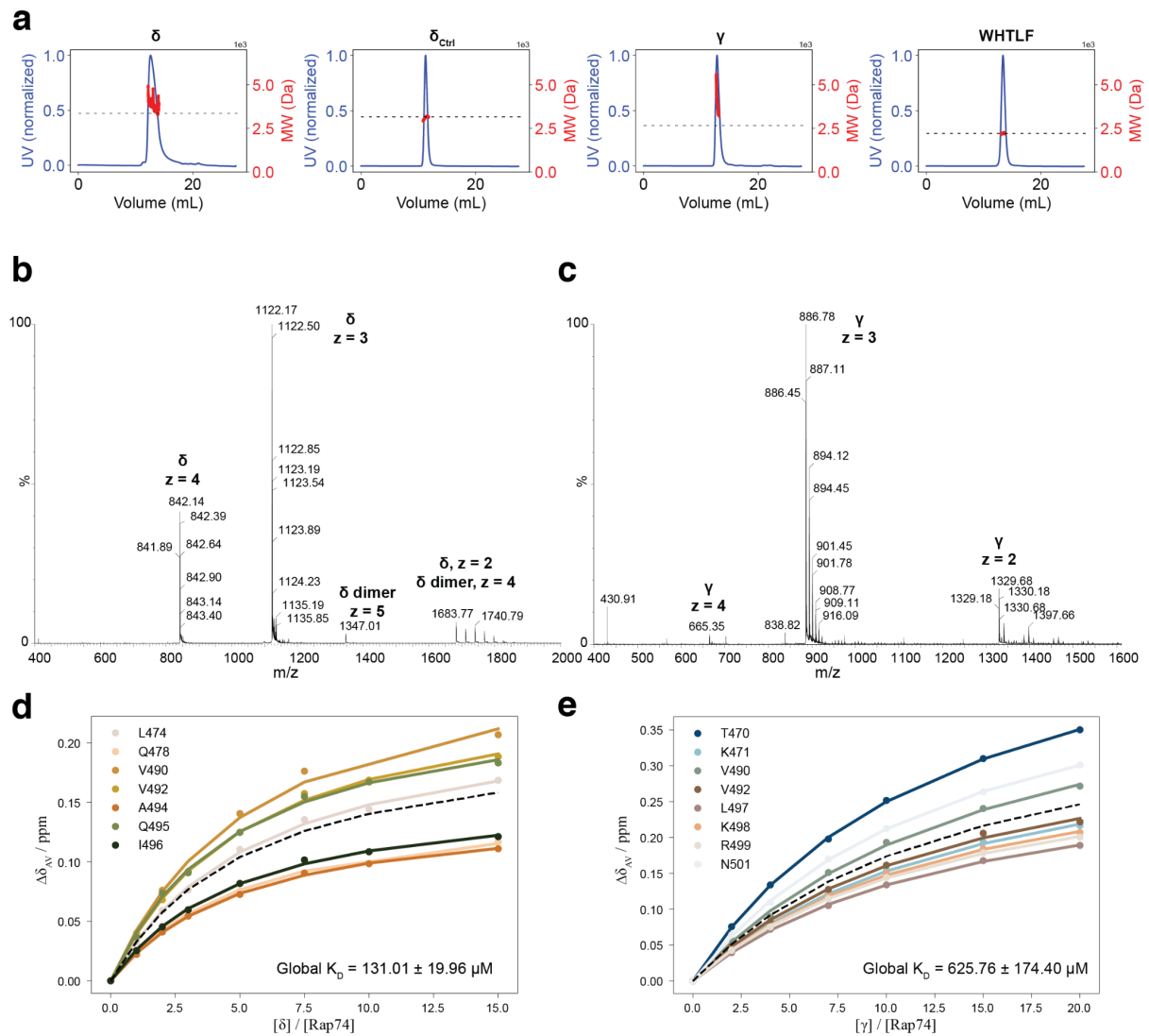


**Fig. S5 | A pH-sensitive conformational switch based on glutamic acid bifurcated hydrogen bonds. a** SEC-MALS analyses of the QXE and E(P3-7)<sub>n</sub> peptides shown in Fig. 4. The horizontal dashed line indicates the molecular weight of the monomeric species calculated using the ProtParam algorithm hosted at ExPASy

(<https://web.expasy.org/protparam/>). **b** Schematic representation of the QM/MM calculations setup. The atoms included in the QM subsystem are shown in sticks and the distances in **c** are indicated with dashed lines. **c** Time series of hydrogen bond donor-acceptor distances throughout four independent QM/MM trajectories. In black, H<sup>N</sup>-O distance. In maroon, H<sub>22</sub>-O distance. Frames in which bifurcated hydrogen bonds are observed are indicated shadowed in gray. **d** Distribution of the distance between the main chain NH of E1 and the main chain CO of L1 in the absence and in the presence of the sc(E1)→mc(L1) hydrogen bond.



**Fig. S6 | SEC-MALS analysis of the TBP polyQ peptide.** Supporting SEC-MALS analysis for the peptide shown in Fig. 5. The horizontal dashed line indicates the molecular weight of the monomeric species calculated using the Protparam algorithm hosted at Expasy (<https://web.expasy.org/protparam/>).



**Fig. S7 | Design of binding motif-grafted peptides to bind the Rap74 domain of TFIIF: supporting analyses.** **a** SEC-MALS analyses of the peptides shown in Fig. 6. The horizontal dashed line indicates the molecular weight of the monomeric species calculated using the ProtParam algorithm hosted at ExPASy (<https://web.expasy.org/protparam/>). **b** Native mass spectra of the peptide  $\delta$  as obtained by Q-TOF MS using a Synapt G1-HDMS (see supplementary methods). Three ionization states corresponding to the monomeric species were independently detected. Two ionization states (one resolved, one ambiguous) corresponding to traces of the dimeric species were detected. **c** Native mass spectra of the peptide  $\gamma$ . Three ionization states corresponding to the monomeric species were independently detected. No oligomeric species were detected. **d**  $K_D$  for the binding of the  $\delta$  peptide to Rap74 as obtained from a global fit (black dashed line) of the binding curves for the top 10% peaks with most intense averaged CSPs (individual fits shown as solid color lines). **e** As in **d** for the binding of the  $\gamma$  peptide to Rap74.

**Table S1 | Details of the peptides used in this study. a** No accompanying symbol indicates that the mass reported was calculated considering  $^{12}\text{C}$  and  $^{14}\text{N}$  isotopes using the ProtParam algorithm hosted at ExPASy (<https://web.expasy.org/protparam/>). † indicates the mass calculated for the  $^{15}\text{N}$  enriched species and ‡ indicates the mass calculated for the  $^{13}\text{C}$ - $^{15}\text{N}$  enriched species. **b** Isotope labeling scheme as in **a**. In addition, no accompanying symbol indicates that the reported mass is an average of that measured during the corresponding peak elution using SEC-MALS, whereas \* indicates that the reported mass is a  $z = 1$  deconvoluted average of the main peaks corresponding to all detected ionization states in the native MS spectra.

| Name                                  | Sequence                          | Expected mass (kDa) <sup>a</sup> | Observed mass (kDa) <sup>b</sup> |
|---------------------------------------|-----------------------------------|----------------------------------|----------------------------------|
| <i>Cooperativity series (Fig. 1)</i>  |                                   |                                  |                                  |
| <b>P1-5</b>                           | KKPGASLLAAQQAQAAAKKY              | 1.92                             | 2.23                             |
| <b>P2-6</b>                           | KKPGASLALAQAQAQAAAKKY             | 1.92                             | 2.19                             |
| <b>P3-7</b>                           | KKPGASLAALQAQAQAAAKKY             | 1.92                             | 2.03                             |
| <b>P5-9</b>                           | KKPGASLAAAQLAAQQAQAAAKKY          | 1.92                             | 2.06                             |
| <b>(P3-7)<sub>2</sub></b>             | KKPGASLAALQALQALQAQAQAAAKKY       | 2.69 <sup>†</sup>                | 2.92 <sup>†</sup>                |
| <b>(P3-7)<sub>3</sub></b>             | KKPGASLAALQALQALQALQALQAQAQAAAKKY | 3.35 <sup>†</sup>                | 3.35 <sup>*†</sup>               |
| <i>L<sub>3</sub>X series (Fig. 3)</i> |                                   |                                  |                                  |
| <b>L<sub>3</sub>WQ<sub>16</sub></b>   | KKPGASLLLLWQQQQQQQQQQQQQQQQQKKY   | 3.79 <sup>†</sup>                | 3.79 <sup>†</sup>                |
| <b>L<sub>4</sub>Q<sub>16</sub></b>    | KKPGASLLLLLQQQQQQQQQQQQQQQQQKKY   | 3.51                             | 3.55                             |
| <b>L<sub>3</sub>YQ<sub>16</sub></b>   | KKPGASLLLLYQQQQQQQQQQQQQQQQQKKY   | 3.76 <sup>†</sup>                | 3.88 <sup>†</sup>                |
| <b>L<sub>3</sub>FQ<sub>16</sub></b>   | KKPGASLLLLFQQQQQQQQQQQQQQQQQKKY   | 3.74 <sup>†</sup>                | 3.76 <sup>†</sup>                |
| <b>L<sub>3</sub>MQ<sub>16</sub></b>   | KKPGASLLLLMQQQQQQQQQQQQQQQQKKY    | 3.73 <sup>†</sup>                | 3.40 <sup>†</sup>                |
| <b>L<sub>3</sub>IQ<sub>16</sub></b>   | KKPGASLLLLIQQQQQQQQQQQQQQQQKKY    | 3.51                             | 3.56                             |
| <b>L<sub>3</sub>EQ<sub>16</sub></b>   | KKPGASLLLLLEQQQQQQQQQQQQQQQKKY    | 3.72 <sup>†</sup>                | 3.61 <sup>†</sup>                |
| <b>L<sub>3</sub>RQ<sub>16</sub></b>   | KKPGASLLLLRQQQQQQQQQQQQQQQKKY     | 3.55                             | 3.42                             |
| <b>L<sub>3</sub>AQ<sub>16</sub></b>   | KKPGASLLLLAQQQQQQQQQQQQQQKKY      | 3.47                             | 3.47                             |
| <b>L<sub>3</sub>KQ<sub>16</sub></b>   | KKPGASLLLLKQQQQQQQQQQQQQKKY       | 3.52                             | 3.45                             |
| <b>L<sub>3</sub>HQ<sub>16</sub></b>   | KKPGASLLLLHQQQQQQQQQQQQQQKKY      | 3.73 <sup>†</sup>                | 3.60 <sup>†</sup>                |
| <b>L<sub>3</sub>VQ<sub>16</sub></b>   | KKPGASLLLLVQQQQQQQQQQQQQKKY       | -                                | -                                |

|                                     |                                 |                   |                   |
|-------------------------------------|---------------------------------|-------------------|-------------------|
| <b>L<sub>3</sub>TQ<sub>16</sub></b> | KKPGASLLLLTQQQQQQQQQQQQQQQQQKKY | 3.55 <sup>†</sup> | 3.40 <sup>†</sup> |
| <b>L<sub>3</sub>SQ<sub>16</sub></b> | KKPGASLLLLSQQQQQQQQQQQQQQQQKKY  | 3.68 <sup>‡</sup> | 3.60 <sup>‡</sup> |
| <i>QXE series (Fig. 4)</i>          |                                 |                   |                   |
| <b>Q1E</b>                          | KKPGASLLLLLEQQQQQQQQQQQQQQQKKY  | 3.51              | 3.62              |
| <b>Q4E</b>                          | KKPGASLLLLQQQEQQQQQQQQQQQQQKKY  | 3.51              | 3.53              |
| <b>Q5E</b>                          | KKPGASLLLLQQQQEQQQQQQQQQQQQKKY  | 3.51              | 3.50              |
| <b>E(P3-7)<sub>2</sub></b>          | KKPGASLAALEALEALEAAEAAKKY       | 2.54              | 2.53              |
| <b>E(P3-7)<sub>3</sub></b>          | KKPGASLAALEALEALEALEALEAAEAAKKY | 3.17              | 3.12              |
| <i>TBP (Fig. 5)</i>                 |                                 |                   |                   |
| <b>TBP</b>                          | KKNSLSILEEQQRQQQQQQQQQQQQQKKY   | 3.66              | 3.67              |
| <i>Grafting series (Fig. 6)</i>     |                                 |                   |                   |
| <b>δ</b>                            | KKPGASLAALQELQLLQLLQELQAAQAACKY | 3.36              | 3.36*             |
| <b>δ<sub>ctrl</sub></b>             | KKPGASAAAAQEAQLAQLAQLAQLAQAACKY | 3.15              | 3.13              |
| <b>γ</b>                            | KKPGASLSWLQLFQEEQAAQAACK        | 2.66              | 2.66*             |
| <b>WHTLF</b>                        | SAAASSSWHTLFTAEEGQLYG           | 2.21              | 2.21              |

## Supplementary Methods

### *Parametrization of $Q_{(i+4)} \rightarrow L_{(i)}$ side chain to main chain hydrogen bonds in Agadir*

Agadir source code was kindly shared by Professor Luis Serrano (Center for Genomic Regulation, Barcelona, Spain). The energy term accounting for  $\text{Leu}_i - \text{Gln}_{i+4}$  interactions  $E_{i,i+4}^L$  was progressively increased in the 0 to 1 kcal mol<sup>-1</sup> range. Chemical shift-derived (BMRB entries 27713, 27714, 27715, 27716, 27717) experimental per-residue helical propensity ( $p_{\text{hel}}$ ) values for the  $L_4Q_n$  peptides (full sequence KKPGASL<sub>4</sub>Q<sub>n</sub>KK) were previously reported<sup>11</sup>. At each  $E_{i,i+4}^L$  increase step, agadir was used to predict per-residue helicity and the RMSD between the predicted and experimentally determined per-residue helical content was obtained using:

$$RMSD_{Hel} = \sqrt{\frac{1}{\text{peptide length}} \sum_{i=1}^{\text{peptide length}} (\text{Hel}\%_i^{\text{experimental}} - \text{Hel}\%_i^{\text{predicted}}(E_{i,i+4}^L))^2}$$

### *SEC-MALS analyses*

The oligomeric state of peptides in solution was determined by size exclusion chromatography coupled to multiple angle light scattering. Peptide samples were loaded in a Superdex Peptide 10/300 GL column (GE Healthcare, Chicago, IL, USA) mounted on a Shimadzu Prominence Modular HPLC with a SPD-20 UV detector (Shimadzu, Kyoto, Japan) coupled to a Dawn Heleos-II multi-angle light scattering detector (18 angles, 658 nm laser beam) and an Optilab T-rEX refractometer (Wyatt Technology, Santa Barbara, CA, USA). The SEC-UV/MALS/RI system was equilibrated with sodium phosphate 20 mM, 0.1% TFA at 298 K with a 0.5 mL min<sup>-1</sup> flow rate and elution was monitored using UV absorbance at 280 nm over 55 minutes. Data acquisition and processing was performed using the Astra 6.1 software (Wyatt Technology, Santa Barbara, CA, USA).

### *Native Mass Spectrometry*

Native mass spectrometry was used to determine the oligomeric state of some peptides. MS experiments were performed using a Synapt G1-HDMS mass spectrometer (Waters, Manchester, UK). All samples were buffer exchanged into 150mM ammonium acetate and were infused by automated chip-based nanoelectrospray using a Triversa Nanomate system (Advion BioSciences, Ithaca, NY, USA) as the interface. The ionization was performed in positive mode using a spray voltage and a gas pressure of 1.75kV and 0.5psi, respectively. The source pumping speed in the backing region (5.85 mbar) of the mass spectrometer was reduced to achieve optimal transmission of non-covalent complexes. Cone voltage,

extraction cone and source temperature were set to 40V, 6V and 40°C, respectively. Trap and transfer collision energies were set to 6V and 4V, respectively. The pressure in the Trap and Transfer T-Wave regions were  $2.39 \cdot 10^{-2}$  mbar of Ar and the pressure in the IMS T-Wave was 0.467 mbar of N<sub>2</sub>. Trap gas and IMS gas flows were 8 and 25 mL/sec, respectively. The instrument was calibrated over the 300-8000Da m/z range using a solution of cesium iodide. MassLynx version 4.1 SCN 704 was used for data processing.