ARTICLES
https://doi.org/10.1038/s41557-020-00626-6
nature chemistry

Predicting the stability of homotrimeric and heterotrimeric collagen helices

Douglas R. Walker[1], Sarah A. H. Hulgan[1], Caroline M. Peterson[1], I-Che Li[1], Kevin J. Gonzalez[1] and Jeffrey D. Hartgerink[1,2] ✉

Robust methods for predicting thermal stabilities of collagen triple helices are critical for understanding natural structure and stability in the collagen family of proteins and also for designing synthetic peptides mimicking these essential proteins. In this work, we determine the relative stability imparted on the collagen triple helix by single amino acids and interactions between amino acid pairs. Using this analysis, we create a comprehensive algorithm, SCEPTTr, for predicting melting temperatures of synthetic triple helices. Critically, our algorithm is compatible with every natural amino acid, can evaluate both homotrimers and heterotrimers, and accounts for all possible helix compositions and registers, including non-canonically staggered helices. We test and optimize our algorithm against 431 published collagen triple helices to demonstrate the quality of our predictive system. Finally, we use this algorithm to successfully guide the design of an ABC heterotrimer possessing high assembly specificity.

Collagen is ubiquitous; it is a vital component of animal proteomes, serves as a foundation of nearly all tissue structures and plays a critical role in the healing of wounds and the behaviour of cancer cells. The characteristic structure of collagen, the triple helix, is found in proteins of the innate immune system such as complement C1q, ficolin and mannose-binding lectins. Other collagen-like proteins are found in viruses, bacteria and fungi[1–4]. Despite their importance, triple helices have been understudied compared to protein structures such as the α-helical coiled-coil. Since the elucidation of the sequence-structure relationship of coiled-coils, beginning with the leucine zipper DNA-binding motif in GCN4 (refs. [5,6]), the study of such protein structures has progressed dramatically, and there are now well-established methods for successful design of impressive arrays of α-helical bundles. Controlled design of α-helical coiled-coils has promoted the understanding of recognition events that are critical in cancer-related pathways such as the Jun–Fos protein pair[7] as well as the production of numerous nanostructures and self-assembling materials[8–14]. By contrast, there has been limited study and exploration of the triple helix structure.

The fundamental structure of collagen, the triple helix, consists of three peptide strands with a polyproline II helical secondary structure that assemble to form a right-handed super helix. Due to steric constraints, every third residue in each strand must be glycine. Between glycines, the positions termed Xaa and Yaa can be occupied by any amino acid, but these are most commonly occupied by proline (Pro) and hydroxyproline (Hyp), respectively[15]. Inter-strand hydrogen bonds between amide protons of glycines and the carbonyl of an amino acid in the Xaa position stabilize the triple helix.

Collagen mimetic peptides (CMPs), which reproduce the triple helical structure, can be designed to have three identical sequences or a combination of different sequences (and are therefore known as homotrimers or heterotrimers, respectively). The preparation of homotrimers is straightforward; for example, $(Pro\text{-}Hyp\text{-}Gly)_n$ ($n \geq 7$) will self-assemble readily. However, the complexity available is greatly restricted; substitutions at the Gly positions lead to catastrophic reductions in helix stability. Substitution of other natural amino acids at the Xaa or Yaa positions reduces the thermal stability of the helix (some unnatural amino acids increase the thermal stability, notably, fluoroproline substitutions at the Yaa position[16]). Additionally, positively charged residues are more stable in the Yaa position, while negatively charged amino acids are more stable in the Xaa position.

The design of heterotrimers, which are biologically common, is much more complicated. Strategies for analysis and design of heterotrimers have been developed only recently[17] and remain inadequate. The main difficulty is the multitude of potential competing species: a mixture of CMPs, A and B, could assemble as A_2B or AB_2, or as a homotrimer of either A or B. As a further complication, steric constraints at the glycine position force a single amino acid offset between each strand. Consequently, each heterotrimer composition exhibits multiple distinct registers; A_2B can assemble as BAA, ABA or AAB, depending on whether B occupies the leading, middle or trailing positions, respectively. For mixtures of three peptides, the number of competing species increases from 8 to 27 (Fig. 1a). A further complication in heterotrimer design is the possibility of non-canonical registrations in which the peptides are offset by more than one amino acid. In such registrations, each cross-section of the triple helix must still contain exactly one glycine to satisfy steric constraints. Non-canonical registers lose the stability conferred by inter-strand hydrogen bonds and van der Waals interactions at their tips; therefore, they melt at lower temperatures than their canonical analogues. However, in some cases, such destabilization has been rectified by stabilizing side-chain interactions[18]. Non-canonical registers can be named according to the offset of each strand in relation to the leading strand; thus, canonically staggered helices are 012 helices, while helices in which the middle strand is offset by an additional triplet are 042 helices. Figure 1b illustrates five offset registers.

Pairwise interactions are critical in the design of heterotrimers as they may either stabilize or destabilize a structure. Given the extended structure of collagen, pairwise interactions are limited to two distinct inter-strand geometries (Fig. 1c,d). Lateral interactions form between amino acids in the Yaa position of the *i*th triplet in one peptide and the Xaa position in the *i*th triplet of the left-handed

[1]Department of Chemistry, Rice University, Houston, TX, USA. [2]Department of Bioengineering, Rice University, Houston, TX, USA. ✉e-mail: jdh@rice.edu

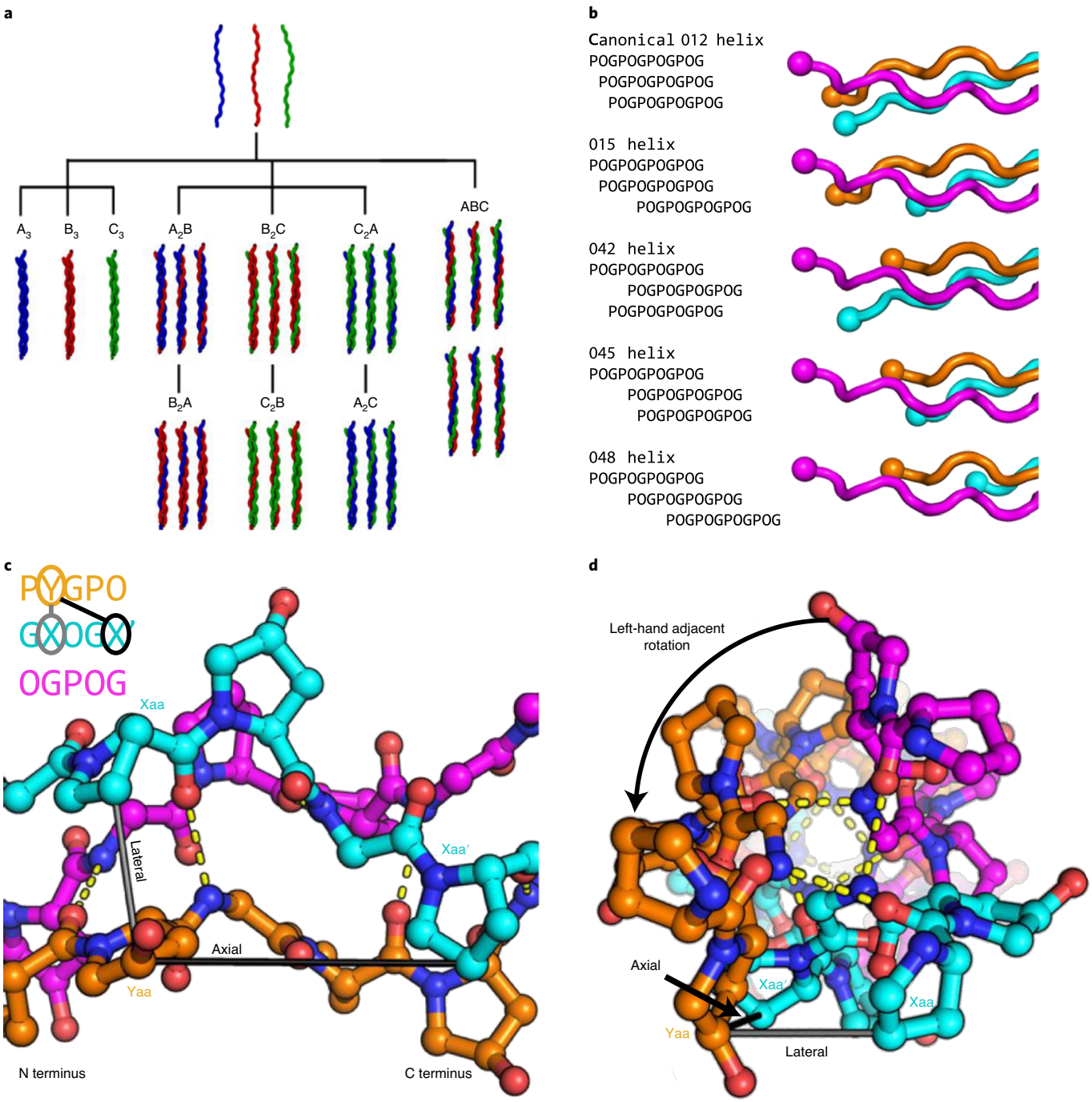


Fig. 1 | Organization of the collagen triple helix. a, The 27 canonical register permutations of three collagen mimetic peptides, A (blue), B (red) and C (green), grouped by composition. A_3, B_3 and C_3 are homotrimers, the central 18 helices are binary heterotrimers and the 6 helices to the right are ternary heterotrimers. **b**, One canonical and four non-canonical arrangements. In all cases, a glycine must reside at every helix cross-section. In canonical 012, the middle and trailing strands are offset by one and two residues, respectively, from the leading strand. In 015, the trailing strand is offset by an additional triplet towards the C terminus compared to the 012 helix. In 042, the middle strand is offset by an additional triplet towards the C terminus. In 045, the middle and trailing strands are both offset by one triplet. In 048, the middle strand is offset by one triplet, and the trailing strand is offset by two triplets. **c**, Text and structure representation of a five-amino-acid-wide section of a collagen helix. N termini are directed to the left. Lassos and bars indicate the possible inter-strand pairwise interactions; lateral interaction (grey bar) between amino acids in the Yaa position (orange) and the Xaa position of an adjacent strand (cyan), and axial interaction (black bar) between amino acids in the Yaa position (orange) and the Xaa′ position (cyan) displaced axially. Backbone hydrogen bonds are indicated by dashed yellow lines. **d,** Collagen helix containing the most common repeat $(Pro\text{-}Hyp\text{-}Gly)_n$, with N termini directed towards the reader. Axial and lateral interactions are approximately parallel and tangential to the helix axis, respectively.

adjacent strand. These residues reside in the same cross-section of the triple helix. Axial interactions form between amino acids in the Yaa position of the *i*th triplet of one peptide and the Xaa position of the $(i+1)$th triplet of the left-handed adjacent strand (Fig. 1c,d). It has been shown that axial interactions between a lysine Yaa residue and a carboxylate Xaa residue can be considerably stabilizing[15,19–22].

Table 1 | Melting temperatures for $(POG)_8$, singly substituted peptides (PYG and XOG) and doubly substituted peptides (XYG and YGX)

Melting temperatures of homotrimers

Xaa\Yaa (T_m °C)		Ala	Phe	His	Lys	Leu	Hyp	Pro	Arg	Val
Ala	XYG						45.0			
	YGX	36.5	25.5			29.0				37.5
Asp	XYG				32.0		42.5	36.5	38.0	
	YGX				47.0				42.0	
Glu	XYG				37.0		44.5	39.5	42.5	
	YGX				44.0				42.5	
Phe	XYG				22.0		37.0	31.0	29.5	
	YGX	29.0	19.5		29.5	24.0			41.5	30.0
Leu	XYG						43.5			
	YGX	34.4	23.5			28.5				34.5
Pro		44.0	31.0	38.0	40.5	36.0	50.0	47.0	48.5	43.5
Val	XYG						41.5			
	YGX	33.5	18.5			19.5				30.5
Trp	XYG			14.5			33.5			
	YGX			22.5						

In contrast, lateral interactions between these residues confer little stabilization[15,21]. These axial interactions are powerful tools in the design of heterotrimeric collagens because isolated lysine, glutamate and aspartate residues are highly destabilizing; this therefore enables both positive design (when charges are paired) and negative design (when unpaired). Other pairwise interactions using charge-pair and cation-π motifs have been studied and developed for heterotrimer design[17,19,23–28]. Many studies have used these interactions to design groundbreaking helices, including the first examples of CMP heterotrimers and various models of diseased collagens[17,18,25,29–33]. However, these systems exhibit the complications of heterotrimer design; most of these helices lack good compositional control and register specificity. Furthermore, despite the utility of these interactions, they are only a subset of the possible pairwise interactions in collagen. There are 380 lateral and 380 axial combinations that may prove useful for triple helix design; hence, a greater understanding of pairwise interactions in collagen is imperative. The use of both stabilizing and destabilizing interactions is critical to optimize specificity in a self-assembled system.

Given the large set of competing species, possible residue substitutions and potential pairwise interactions, the advantages of a computational approach to predict triple helix stability are apparent. Such an approach can rapidly assess the stabilities of all competing species and the specificity with which peptides will self-assemble into the desired helix. In 2005, Persikov et al. published the first algorithm to predict the stability of CMPs, which has been the gold standard tool to aid CMP design[15]. This program, hereafter termed P-CSC (Persikov's Collagen Stability Calculator), uses the relative propensities of all natural amino acids in the Xaa and Yaa positions and considers select amino acid interactions for the stabilization of triple helices. However, P-CSC assesses only homotrimers. Subsequently, many other algorithms to assess and design heterotrimeric collagens have been published[23,25,30,31,34,35]; however, none, including P-CSC, have been able to predict melting temperatures with sufficient precision. An in-depth discussion of the algorithmic precedents to our study is included in the Supplementary Discussion (see the section 'Algorithmic precedents').

Algorithms that simultaneously predict stability, composition, register and specificity for collagen triple helices could play a key role in the evaluation of natural collagens and collagen-like proteins and in the de novo design of peptide systems. Here, we describe SCEPTTr (Scoring function for Collagen-Emulating-Peptides' Temperature of Transition), an algorithm that makes such predictions with high accuracy and precision (see Supplementary Software 1). The performance of SCEPTTr is tested against 431 published triple helices and is then used to design an ABC heterotrimer using amino acids that have not yet been combined in heterotrimer design. We experimentally verify the thermal stability, specificity, composition and register of this triple helix and demonstrate that these parameters show a close match with the predictions of SCEPTTr.

Results and discussion

To increase the number of pairwise interactions that are understood, we synthesized 49 peptides to analyse 33 of the 760 possible interactions. The peptides selected were chosen primarily for two reasons. First, peptides that have been described previously in the literature had been characterized using a variety of experimental methods that can influence the observed melting temperature. We wanted to use a unified method to repeat the determination of the melting temperature for what we expected to be the most important

Table 2 | Results from the algebraic deconvolution, showing the effect of each axial or lateral pairwise interaction on the stability of the triple helix

Calculated stabilizing or destabilizing effects conferred by axial and lateral interactions

Xaa\Yaa (Δ)		Ala	Phe	His	Lys	Leu	Hyp	Pro	Arg	Val
Ala	Lateral									
	Axial	–1.250	–0.250			–1.000				–0.500
Asp	Lateral				–0.500			–1.500	–1.500	
	Axial				7.250				1.250	
Glu	Lateral				1.000			–1.000	–0.250	
	Axial				4.000				–0.125	
Phe	Lateral				–2.750			–1.500	–3.000	
	Axial	–1.000	0.750		2.375	0.500			4.500	–0.250
Leu	Lateral									
	Axial	–1.550	–0.500			–0.500				–1.250
Pro										
Val	Lateral									
	Axial	–1.000	–2.000			–4.000				–2.250
Trp	Lateral			–3.500						
	Axial			2.250						

The intensity of the colour gradient correlates with the magnitude of stabilizing interactions (in cyan) and destabilizing interactions (in orange). The values have units of °C per interaction and indicate that each interaction was found to stabilize (positive value) or destabilize (negative value) a triple helix by the number of degrees specified.

amino acid substitutions. Second, specific amino acid combinations that have not been described previously in the literature were added due to their perceived importance in natural collagens.

A single substitution in the peptide $(POG)_n$ yields a homotrimer with three substituted residues present (Supplementary Fig. 1). Owing to their triple helical geometry, these residues extend radially and cannot interact with one another. Therefore, the thermal effect of that substitution is due primarily to the inherent propensity of that residue to mediate helix stability. However, double substitutions become more complicated, and two cases, XYG and YGX, need to be considered, as described in the Supplementary Discussion (see the section 'Deconvolution of axial and lateral interactions').

As outlined in the Methods section, peptides were synthesized to ascertain the effects of amino acid combinations on the stability of a collagen triple helix (Table 1). Table 1 shows 49 homotrimer melting temperatures. Circular dichroism melting curves and their derivatives are available in Supplementary Figs. 2–50, as are the intermediate values for deconvoluting interactions from melting temperatures (Supplementary Table 1). Table 2 shows the calculated thermal stability contributions of axial and lateral interactions in °C per interaction. Notably, in agreement with the literature[15,24], the strongest interaction is the axial lysine-aspartate interaction ($Lys\text{-}Asp_{AX}$) with a stabilization of 7.3 °C per interaction. Other notable stabilizing pairwise interactions include $Arg\text{-}Phe_{AX}$ (4.5 °C), $Lys\text{-}Glu_{AX}$ (4.0 °C), $Lys\text{-}Phe_{AX}$ (2.4 °C) and $His\text{-}Trp_{AX}$ (2.3 °C). Notable destabilizing interactions include $Leu\text{-}Val_{AX}$ (−4.0 °C), $His\text{-}Trp_{LAT}$ (−3.5 °C), $Arg\text{-}Phe_{LAT}$ (−3.0 °C), $Lys\text{-}Phe_{LAT}$ (−2.8 °C), $Val\text{-}Val_{AX}$ (−2.3 °C) and $Phe\text{-}Val_{AX}$ (−2.0 °C).

Computational design. The above analysis of pairwise interactions was used in the design of SCEPTTr. The algorithm was tested against a library of 431 published triple helices. SCEPTTr considers six classes of variables to determine the thermal stability of a collagen helix. These classes and the corresponding sources for the values used are discussed next. First, in terms of length, melting temperatures of $(GPO)_n$ homotrimers of various lengths have been obtained from previous studies[15,17,36]. The $(POG)_9$ homotrimer has not been characterized in the literature. Therefore, the melting temperature for $(POG)_9$ was calculated by applying the relationship between temperature and length for the $(GPO)_n$ series to the $(POG)_n$ series. Second, terminal functionality has been shown to affect the stability of triple helices. In SCEPTTr, the values for stabilization resulting from N-terminal and C-terminal functionalization have been derived from Egli et al.[36] Third, considering amino acid propensity, values for additional thermal destabilizations of single amino acids come from Persikov et al.[15] Fourth, for pairwise interactions, in addition to those calculated in the previous section, pairwise interactions were determined by and acquired from refs. [19,24,27,37]. Fifth, for triplet-dependent weighting, a previous study[38] found that the tips of a triple helix fray before the core in a melting analysis. Therefore, triplets nearer the N and C termini of peptides in a triple helix have a lower impact on its overall stability. This difference in stability from tips to core is accounted for by the triplet-dependent weighting function in SCEPTTr. Last, considering registration, canonical registers are evaluated independently, followed by all non-canonical registers. Non-canonical

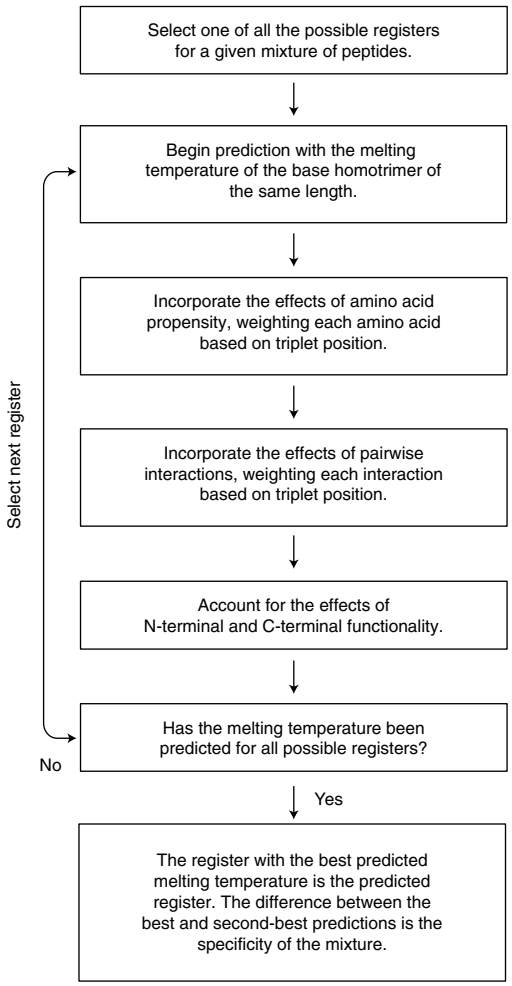


Fig. 2 | Flow chart of the logical flow of SCEPTTr to predict melting temperatures of combinations of peptides in collagen-like triple helices. Given a selection of CMPs of a particular triple helix, SCEPTTr begins at the first step (top) in this flow chart and selects one of the possible registers. Then, SCEPTTr considers the lengths of the peptides, the amino acids contained in the triple helix, the pairwise interactions present in the triple helix and the functionality of the termini of the peptides to calculate the melting temperature for the given register. This process is repeated for all of the registers. Next, the results are sorted by melting temperature to display the predicted melting temperature of the combination of CMPs and the specificity of the system.

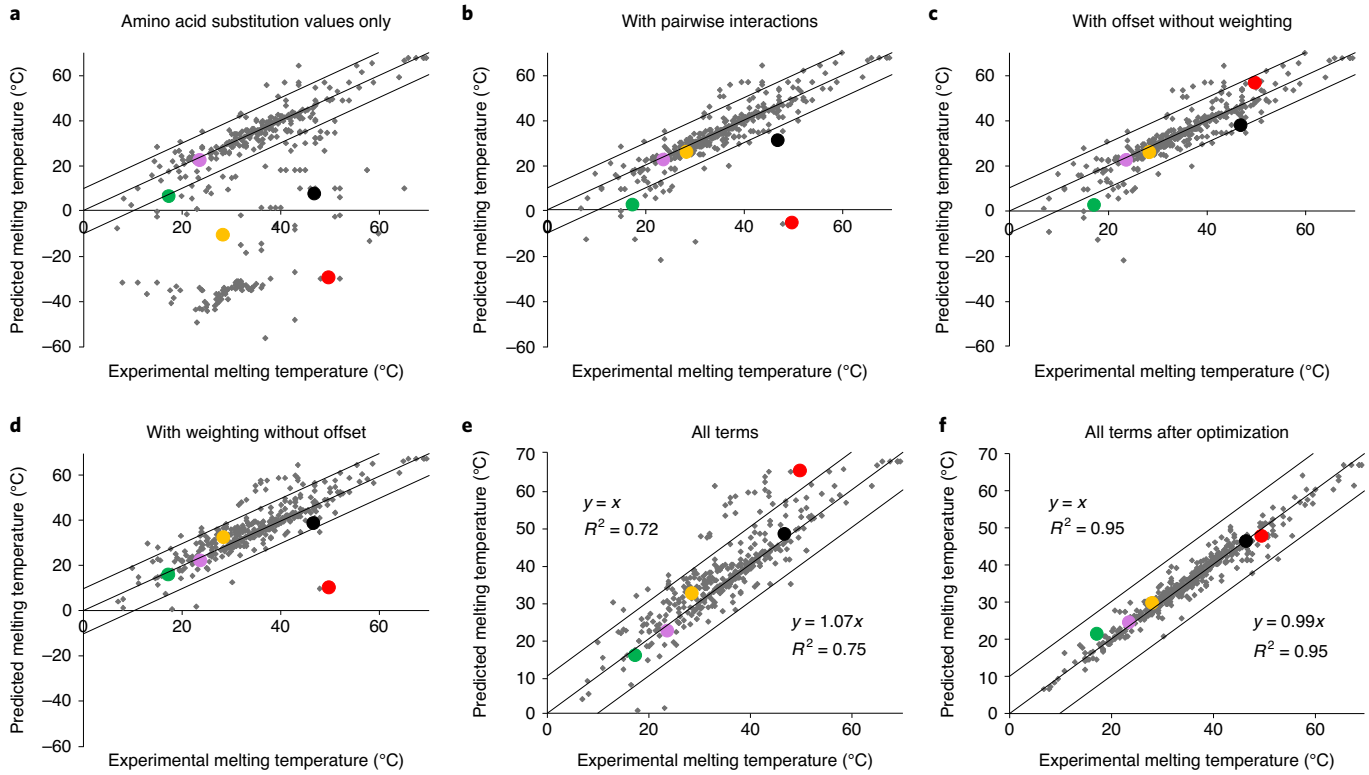
registers are destabilized based on melting temperatures reported in ref. [18]. SCEPTTr will consider all registers that have a maximum offset of three triplets. Thus, 9 registers are evaluated for homotrimers, 72 registers for binary heterotrimers and 243 registers for ternary heterotrimers. A complete list of parameters used by SCEPTTr can be found in Supplementary Table 2. The flow chart for SCEPTTr is shown in Supplementary Fig. 59.

Example to illustrate the scoring function process. The scoring function of SCEPTTr can be demonstrated by considering the flow chart in Fig. 2 and the homotrimer formed by GARGPAGPQGPRGDKGETGPOGPOGPOGPOGV[15] as an example. First, the helix is assigned a melting temperature based on the length of the peptides. Many synthetic peptides are synthesized with extra amino acids, herein termed 'tags', that break the triplet repeat at one of the termini for practical reasons. SCEPTTr has no understanding of the effect of tags, and therefore, we remove them before performing the analysis. In this example, we disregard the C-terminal –GV tag in the sequence above. $(GPO)_{10}$ is considered to be the parent sequence, which melts at 63.8 °C. Next, SCEPTTr considers the amino acid substitutions necessary to convert the parent sequence to the sequence of interest. Each substitution has an associated destabilizing value, which can be adjusted by the weighting function. The substitutions in the example result in a 14.0 °C penalization for each peptide strand in the helix. The next consideration is the weighted pairwise interactions. The canonical homotrimer contains five axial interactions (two K-E, two R-D and one R-E) resulting in a 10.0 °C increase in helix stability. When these are considered together, the estimated melting temperature can be calculated to be 31.7 °C, a good estimate of the reported melting temperature of 30.8 °C (ref. [15]). SCEPTTr performs the same analysis for all of the other relevant offset registers. The melting temperature is adjusted according to the stagger of the considered species; the most stable species for this peptide is canonical and incurs no penalty. The second-best species is a 015 stagger and incurs a penalty of −3.6 °C corresponding to lost backbone H-bonds at the tips of the helix. Then, SCEPTTr compares the estimated melting temperatures of all registers to determine the most stable arrangements. The specificity of the system can then be calculated as the difference between the two highest melting temperatures.

Performance evaluation of SCEPTTr. To evaluate the performance of SCEPTTr for a wide range of sequences, a library containing 431 triple helices was assembled; it included a sampling of all natural amino acids, homotrimers and both A_2B and ABC heterotrimers, and contained sequences from many different research groups. The sequences of peptides in this library and their corresponding references are provided in Supplementary Table 5. SCEPTTr analysed all of the 431 helices. The published versus predicted melting temperatures are plotted in Fig. 3e. The helix composition predicted by SCEPTTr was constrained to match the result published, but the register was allowed to vary. Given that, in the ideal case, the predicted T_m will match published values exactly, the line of best fit for any estimation should approach $y = x$. As seen in Fig. 3e, the line through the SCEPTTr predictions, $y = 1.07x$, is near ideal.

To evaluate the terms used by SCEPTTr, each was removed and subsequently reincorporated stepwise. To compare the quality of each prediction, we calculated the residual sum of squares, $SS_{res} = \sum(T_{m,pred} - T_{m,meas})^2$. We report these as quotients of the library size to contextualize the values (SS_{res}/n). Figure 3a illustrates the predictions when SCEPTTr understands amino acid substitution effects, peptide length, and amidated and/or acetylated termini. With no additional terms, SCEPTTr accurately predicts only a subset of the analysed triple helices (helix 1 of Fig. 3g), but produces abysmal estimations for helices with a higher degree of complexity, as evident from $SS_{res}/n = 1{,}105.1$. For comparison, an SS_{res}/n value of 100 indicates an average error of 10 °C. Figure 3b illustrates the analysis by SCEPTTr when the knowledge of pairwise interactions is available. The fit is improved for many species, as shown by helices 2–6 and the dramatic decrease of SS_{res}/n to 72.2. Figure 3c illustrates the analysis by SCEPTTr when the additional allowance for non-canonical offsets without tip-to-core weighting is available. The accuracy improved for a subset that SCEPTTr predicts as being stabilized by non-canonical registers (helices 3 and 5), resulting in an SS_{res}/n value of 51.9. In Fig. 3d, SCEPTTr uses the weighting function but disallows non-canonical registers, lowering SS_{res}/n to 47.7. This result indicates that both the weighting function and non-canonical registers are important for accurate prediction of triple helix melting temperatures. Finally, the combination of all of the terms produces the best model, with $SS_{res}/n = 45.6$ (Fig. 3e).

Optimization by genetic algorithm. The parameters of SCEPTTr were then optimized by a genetic algorithm. A complete discussion



g

Helix		Peptides	Prefix	Suffix	Capped	T_m	a	b	c	d	e	f
1	●	GPOGPOGROGKRGKQGQKGPOGPOGPOGPO	–	G	Y	23.9[40]	21.90	21.90	21.90	21.90	21.90	24.26
2	●	PKGPKGPKGKOGPDGAOGDOGDOGPKGPKG PDGDOGDOGDOGPDGKOGPDGPDGPDGDOG KOGPDGPDGPKGKOGPKGKOGKOGKOGKOG	YG	–	N	28.6[30]	–11.10	25.32	25.32	31.88	31.88	29.48
3	●	DKGPOGDKGPOGDKGPOGDKGPOGDKGPOG	–	–	N	50.0[39]	–30.25	–6.24	56.01	9.95	64.74	47.41
4	●	GLAGAOGLRGGAGPOGPEGPOGPOGPOGPO	–	GV	Ac	17.5[41]	5.90	1.85	1.85	15.41	15.41	21.35
5	●	POGPKGPOGPKGPOGPKGPOGPKGPOGPKG EOGPOGEOGPOGEOGPOGEOGPOGEOGPOG PKGEOGPKGEOGPKGEOGPKGEOGPKGEOG	WG	–	Am	47.0[18]	7.00	30.67	37.67	38.33	47.63	46.05

Fig. 3 | Performance of SCEPTTr for the library of triple helices available in the literature. a–f, Results of the analysis by SCEPTTr for amino acid propensities only (**a**), added pairwise interactions (**b**), added potential for offset registrations without the weighting function (**c**), added weighting function without potential for offset registrations (**d**), all terms together (**e**), and all terms after optimization by genetic algorithm (**f**). The three solid lines in each plot are the lines $y = x$, $y = x + 10$ and $y = x - 10$. **g**, Data points emphasized in colour in (**a**)–(**f**) are shown along with their experimental (column T_m) and predicted melting temperatures (columns a–f). Values given in columns a–f correspond to the points plotted in (**a**)–(**f**). 'Prefix' and 'suffix' refer to tags at the front and end of a sequence, respectively. 'Capped' indicates whether peptides are amidated (Am), acetylated (Ac), neither (N) or both (Y). The coloured symbols correspond to the markers[18,30,39–41] emphasized in the plots.

of this process can be found in the Supplementary Discussion (see the section 'Optimization of SCEPTTr') and in a flow chart (Supplementary Fig. 57). This algorithm was allowed to adjust the parameters of SCEPTTr for any pairwise interactions observed. Before the genetic algorithm was incorporated, SCEPTTr used 27 experimentally defined pairwise interactions. When the genetic algorithm was used, 103 useful interactions were employed, the R^2 value for the prediction fit increased to 0.95 (Fig. 3f) and SS_{res}/n decreased to 5.95. These new interactions should be verified experimentally due to their exciting potential. Furthermore, it was discovered during optimization that additional features are important for broader and more accurate predictions; these features include a more generalized evaluation of peptide length and N-terminal and C-terminal functionality, along with incorporation of localized charge density and adjacent Xaa, Yaa substitutions. A discussion of the inclusion of these features can be found in the Supplementary Discussion (see the section 'New features found through optimization'). Supplementary Table 3 shows the values used by SCEPTTr after optimization. Other limitations of SCEPTTr are discussed in the Supplementary Discussion (see the section 'Weaknesses of the model'). A statistical discussion of the performance of SCEPTTr can also be found in the Supplementary Discussion (see the section

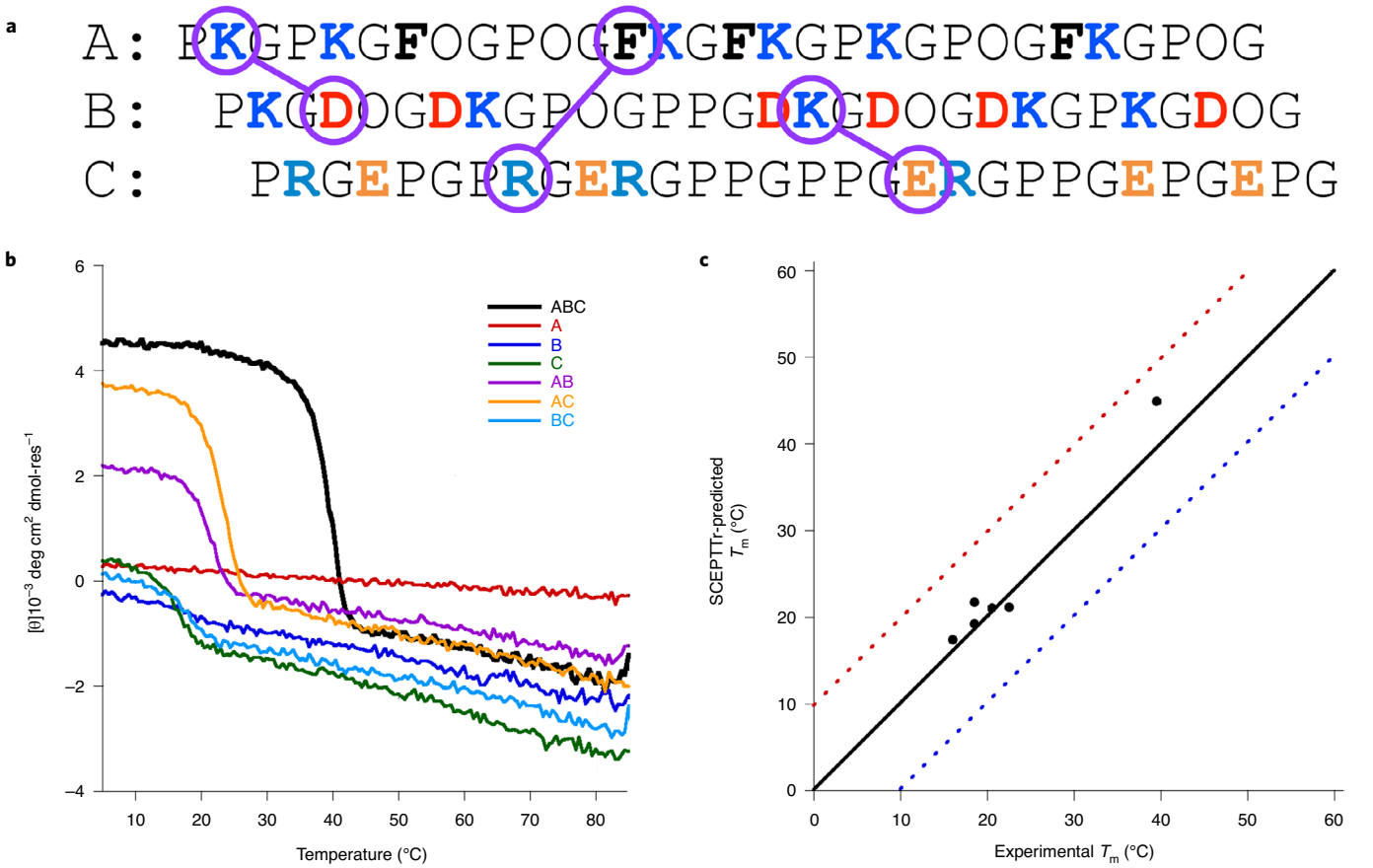


Fig. 4 | Thermal characterization of the ABC heterotrimer. a, Sequences of the three peptides that constitute the ABC heterotrimer and their expected registration. In each peptide, glycine #21 is ^{15}N-labelled while N-termini are acetylated and C-termini are amidated. Purple lassos indicate axial interactions between Lys-Asp, Arg-Phe and Lys-Glu that stabilize the indicated composition and register. For clarity, only 3 of 15 such interactions are highlighted. **b**, Thermal melting data from circular dichroism for the ABC heterotrimer and each of its components. The ABC heterotrimer (black) is clearly seen to have the highest T_m. **c**, Comparison of the experimentally determined T_m and the T_m predicted by SCEPTTr for each component with an observable circular dichroism transition. Dotted lines indicate 10 degrees over (red) or under (blue) the estimate.

'Statistical analysis'). Despite its existing limitations, SCEPTTr predicts the melting temperature of collagen-mimetic peptide triple helices with precision, as observed in Fig. 3f.

Use of SCEPTTr in heterotrimer design. Before its final optimization, SCEPTTr was used to aid the design of an ABC heterotrimeric triple helix consisting of the peptides shown in Fig. 4a. It is important to note that these peptides were designed by humans, but this design process was expedited by the assistance of SCEPTTr. The use of SCEPTTr enabled the implementation of many small changes to the sequence during design and a rapid analysis of the effects of those changes, calculating alterations in both stability and specificity in seconds. For example, each of the prolines in Yaa positions are used to destabilize competing compositions, which may not be obvious without a comprehensive examination of all competing species. The C peptide homotrimer was particularly difficult to destabilize and therefore required substitution of all hydroxyprolines. By contrast, although the B peptide homotrimer also required destabilization, many hydroxyproline residues could be retained. Furthermore, the placement of each amino acid and each axial interaction was honed by combining human intuition with computational assessment, optimizing positive interactions in the desired species and negative interactions in competing species. Our ABC triple helix contains stabilizing Lys-Asp, Lys-Glu and Arg-Phe axial interactions that are highlighted in Fig. 4a; however, the design also takes into account destabilizing Lys-Phe and Arg-Phe lateral interactions to sabotage competing species. These lateral destabilizing interactions are not apparent without thorough analysis performed by a computational algorithm such as SCEPTTr. Therefore, a process that takes a day when SCEPTTr is used for assistance would require weeks if the same thorough process to assess each competing species were performed by hand. The resulting ABC heterotrimer consists of the most diverse combination of amino acids used in the design of a triple helix until now.

Experimental verification of the ABC triple helix. Circular dichroism melting analysis was performed on each peptide alone, each binary mixture at 1:1 ratios and the ternary mixture in a 1:1:1 ratio. The results of these experiments are presented in Fig. 4b. From these data, we observe that the ternary mixture results in a unique species with higher thermal stability (39.5 °C) than any competing species. The highest T_m of the competing species (23.5 °C) can then be used to calculate the specificity of the system, which is 16 °C. These values are plotted against predictions from SCEPTTr in Fig. 4c, demonstrating that the entire ABC system is well predicted.

Although circular dichroism is important for defining thermal stability, it is limited in its ability to evaluate the composition and register of mixed peptide systems. To address this limitation, we used NMR. Each of the three peptides was synthesized with one ^{15}N-labelled glycine at position 21. This is a powerful method to

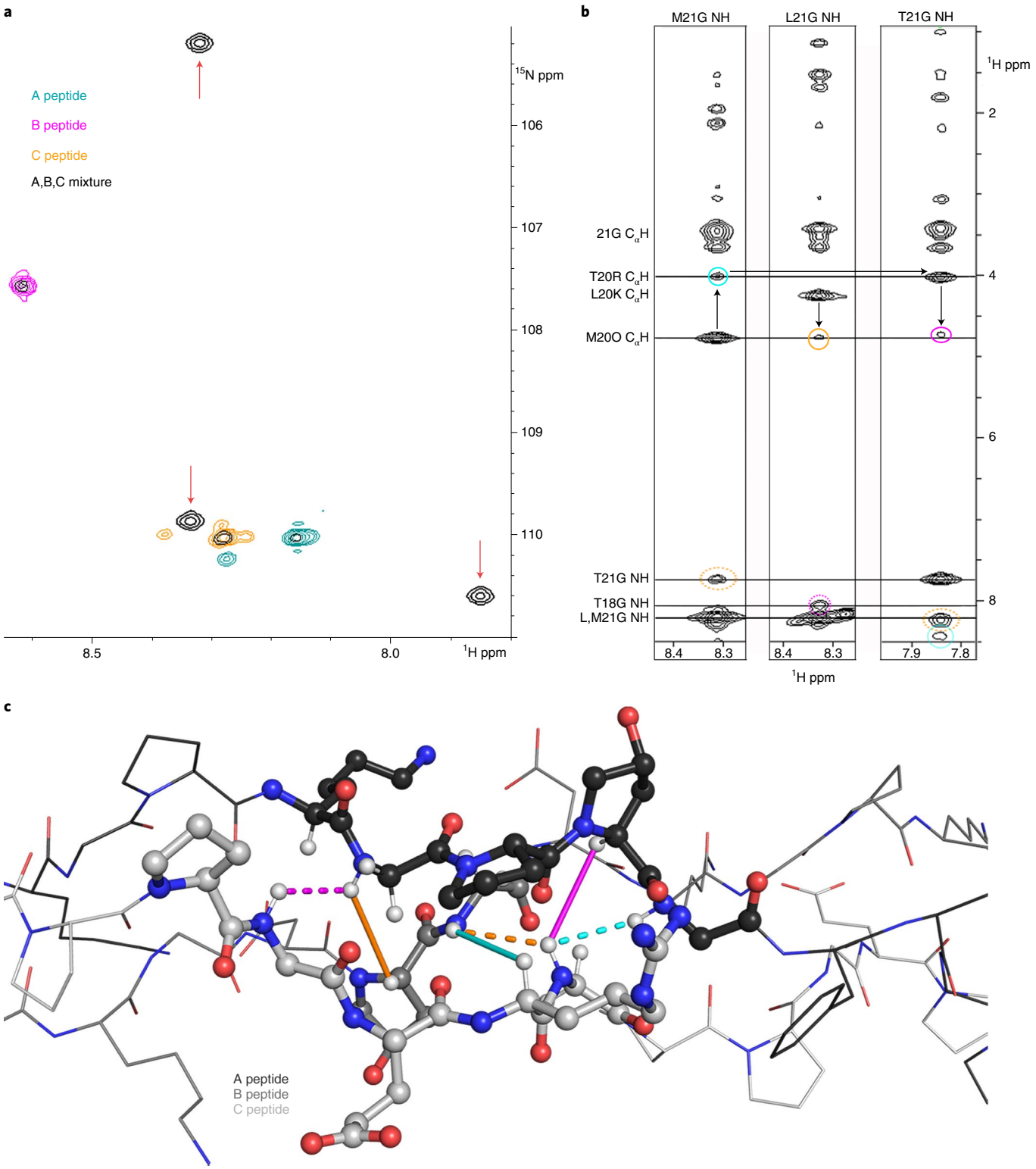


Fig. 5 | Structural characterization of the ABC heterotrimer. a, ^{1}H–^{15}N HSQC of the unary and ternary peptide solutions. The peaks for A, B and C peptide solutions alone indicate that they contain almost entirely monomeric species. Black peaks correspond to the ternary peptide solution. The peaks that overlap between the unary and ternary samples indicate monomeric peptides, while the three unique peaks in the ternary sample indicate triple helices (red arrows). The absence of other peaks in the HSQC indicates the formation of a single-composition, single-register triple helix in solution. **b**, Three NOESY planes taken from a 3D ^{1}H–^{1}H–^{15}N NOESY HSQC at specific ^{15}N shifts correspond to amide trimer peaks in HSQC; from left to right, 105.2 ppm, 109.8 ppm and 110.5 ppm. Side chain analysis (Supplementary Discussion, see the section 'Discussion of NMR peak assignments') shows that these amides belong to the B peptide, the A peptide and the C peptide, respectively. Circled cross-peaks and the progression of the arrow indicate through-space interactions and prove that the in situ register matches the predicted register. Atom labels are included at the left side of the spectra to indicate important proton shifts. **c**, Molecular model of the triple helix. The circle colours and styles in (**b**) match the interaction lines in (**c**) as follows: dotted magenta indicates leading Gly to unlabelled trailing Gly NOE; dotted orange indicates middle Gly to trailing Gly; dotted cyan indicates trailing Gly to unlabelled leading Gly; solid orange indicates leading Gly to middle Hyp; solid cyan indicates middle Gly to trailing Arg; and solid magenta indicates trailing Gly to leading Hyp.

simplify complicated and highly overlapping NMR spectra. $^{1}H–^{15}N$ heteronuclear single quantum coherence (HSQC) experiments will show a single peak for each chemically distinct conformation of each peptide. Further explanation of the interpretation of HSQC experiments can be found in the Methods. Figure 5a shows the overlay of four NMR experiments: three from each peptide alone and one from the ternary mixture at a ratio of 1:1:1. (Three additional experiments for the binary mixtures at 1:1 ratios are presented in Supplementary Fig. 55.) The plot in Fig. 5a shows that each single peptide adopts primarily one unfolded, monomeric conformation, except C, which forms a small fraction of homotrimer, consistent with circular dichroism. In contrast, the ABC mixture displays three additional peaks (indicated with red arrows) that are not observed in any single peptide or double peptide mixture. These three peaks are indicative of the formation of a triple helix with the composition of one A, one B and one C peptide. It also strongly suggests that a single registration of triple helix is present.

However, to determine the register without ambiguity, we used a 3D nuclear Overhauser effect spectroscopy (NOESY) HSQC NMR experiment. A more in-depth explanation of the interpretation of NOESY HSQC experiments can be found in the Methods. Figure 5c and Supplementary Video 1 are both useful to illustrate the following discussion. Figure 5b shows three NOESY planes of that experiment taken at ^{15}N shifts corresponding to the triple helical peaks seen in the HSQC (Fig. 5a) for 105.2 ppm, 109.9 ppm and 110.5 ppm. Each vertical line of peaks arises from through-space NOE interactions between a labelled amide proton and other nearby protons. The large Cα-H peaks (~3.5–5 ppm) indicate intra-strand NOE interactions with Yaa amino acids preceding each glycine. The small Cα-H peaks (~3.5–5 ppm) are inter-strand correlations between labelled glycine amide N-H and Cα-H on the succeeding peptide chain in the registration. Starting at a large Cα-H peak and moving vertically to a small Cα-H peak and horizontally to the next Cα-H results in the registration of the triple helix from leading to middle to trailing strands as indicated by the arrows in Fig. 5b. The more upfield peaks inform strand identity and are discussed in the Supplementary Discussion (see the section 'Discussion of NMR peak assignments'). Note that the trailing strand amide N-H shows a small NOE to an otherwise unseen Cα-H (circled in magenta in the NOESY plane that is farthest right in Fig. 5b) corresponding to a position 23 hydroxyproline. Figure 5c shows the positions of these interactions within the triple helix and demonstrates that these correlations can only be observed when the triple helix is in the register predicted by SCEPTTr; A leading, B middle, C trailing. A more detailed discussion of these data can be found in the Supplementary Discussion (see the section 'Discussion of NMR registration assignment'). The collective analysis of circular dichroism and NMR show a melting temperature of 39.5 °C, a specificity of 16 °C and a preferred ABC registration, all of which are predicted accurately by SCEPTTr.

Conclusion

We have presented a robust method to elucidate the thermal effects of amino acid pairwise interactions on collagen triple helices and have applied the method to a set of 49 homotrimers to deconvolute the effect of 33 pairwise interactions on stability. These effects, along with other values from the literature, were used to create a comprehensive algorithm, SCEPTTr, to predict the melting temperatures of CMPs. Importantly, SCEPTTr can successfully evaluate homotrimers and A_2B and ABC heterotrimers. Results obtained from testing SCEPTTr against a library of 431 published collagen triple helices demonstrate its ability to accurately predict melting temperatures with precision. SCEPTTr is unique due to the combination of five features: (1) it can assess peptides with any of the 20 naturally expressed amino acids as well as hydroxyproline; (2) it can assess peptide mixtures with any combination of collagen mimetic peptides, both homotrimers and heterotrimers; (3) it assesses all possible combinations (different compositions and registers) of the peptides that it is given and predicts an experimentally testable melting temperature (T_m) for each; (4) through this assessment, it enables the prediction of the experimentally observable triple helix composition and register of the given set of peptides; (5) and it enables the prediction of the specificity of the given set of peptides. The specificity indicates whether the most stable species will assemble to the exclusion of all others or whether a complex mixture of triple helices will be formed. The combination of these five features and the accuracy of prediction by SCEPTTr distinguishes this algorithm from all other collagen predictive algorithms. Finally, we have shown that SCEPTTr can be used to achieve objectives that previous algorithms could not: to inform discoveries of more pairwise interactions and other parameters affecting collagen stability and to enable the design of novel heterotrimers with high degrees of specificity. We believe that this algorithm will help to advance the collagen field towards achieving the versatility of design of α-helices and the range of specificity available to DNA complementarity.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41557-020-00626-6.

Received: 30 August 2019; Accepted: 14 December 2020;
Published online: 15 February 2021

References

1. Celerin, M. et al. Fungal fimbriae are composed of collagen. *EMBO J.* **15**, 4445–4453 (1996).
2. Rasmussen, M., Jacobsson, M. & Björck, L. Genome-based identification and analysis of collagen-related structural motifs in bacterial and viral proteins. *J. Biol. Chem.* **278**, 32313–32316 (2003).
3. Thiel, S. Complement activating soluble pattern recognition molecules with collagen-like regions, mannan-binding lectin, ficolins and associated proteins. *Mol. Immunol.* **44**, 3875–3888 (2007).
4. Yu, Z., An, B., Ramshaw, J. A. M. & Brodsky, B. Bacterial collagen-like proteins that form triple-helical structures. *J. Struct. Biol.* **186**, 451–461 (2014).
5. O'Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science* **254**, 539–544 (1991).
6. Landschulz, W. H., Johnson, P. F. & McKnight, S. L. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science* **240**, 1759–1764 (1988).
7. O'Shea, E. K., Rutkowski, R., Stafford, W. F. III & Kim, P. S. Preferential heterodimer formation by isolated leucine zippers from fos and jun. *Science* **245**, 646–648 (1989).
8. Lovejoy, B. et al. Crystal structure of a synthetic triple-stranded α-helical bundle. *Science* **259**, 1288–1293 (1993).
9. Zimenkov, Y., Conticello, V. P., Guo, L. & Thiyagarajan, P. Rational design of a nanoscale helical scaffold derived from self-assembly of a dimeric coiled coil motif. *Tetrahedron* **60**, 7237–7246 (2004).
10. Zimenkov, Y. et al. Rational design of a reversible pH-responsive switch for peptide self-assembly. *J. Am. Chem. Soc.* **128**, 6770–6771 (2006).
11. Moutevelis, E. & Woolfson, D. N. A periodic table of coiled-coil protein structures. *J. Mol. Biol.* **385**, 726–732 (2009).
12. Fallas, J. A. et al. Computational design of self-assembling cyclic protein homo-oligomers. *Nat. Chem.* **9**, 353–360 (2017).
13. Rhys, G. G. et al. Maintaining and breaking symmetry in homomeric coiled-coil assemblies. *Nat. Commun.* **9**, 4132 (2018).
14. Koepnick, B. et al. De novo protein design by citizen scientists. *Nature* **570**, 390–394 (2019).
15. Persikov, A. V., Ramshaw, J. A. M. & Brodsky, B. Prediction of collagen stability from amino acid sequence. *J. Biol. Chem.* **280**, 19343–19349 (2005).
16. Holmgren, S. K., Bretscher, L. E., Taylor, K. M. & Raines, R. T. A hyperstable collagen mimic. *Chem. Biol.* **6**, 63–70 (1999).
17. Gauba, V. & Hartgerink, J. D. Self-Assembled heterotrimeric collagen triple helices directed through electrostatic interactions. *J. Am. Chem. Soc.* **129**, 2683–2690 (2007).

18. Jalan, A. A., Jochim, K. A. & Hartgerink, J. D. Rational design of a non-canonical "sticky-ended" collagen triple helix. *J. Am. Chem. Soc.* **136**, 7535–7538 (2014).
19. Persikov, A. V., Ramshaw, J. A. M., Kirkpatrick, A. & Brodsky, B. Peptide investigations of pairwise interactions in the collagen triple-helix. *J. Mol. Biol.* **316**, 385–394 (2002).
20. Persikov, A. V., Ramshaw, J. A. M., Kirkpatrick, A. & Brodsky, B. Electrostatic interactions involving lysine make major contributions to collagen triple-helix stability. *Biochemistry* **44**, 1414–1422 (2005).
21. Fallas, J. A., Dong, J., Tao, Y. J. & Hartgerink, J. D. Structural insights into charge pair interactions in triple helical collagen-like proteins. *J. Biol. Chem.* **287**, 8039–8047 (2012).
22. Wei, F., Fallas, J. A. & Hartgerink, J. D. Sequence position and side chain length dependence of charge pair interactions in collagen triple helices. *Macromol. Rapid Commun.* **33**, 1445–1452 (2012).
23. Xu, F., Zhang, L., Koder, R. L. & Nanda, V. De novo self-assembling collagen heterotrimers using explicit positive and negative design. *Biochemistry* **49**, 2307–2316 (2010).
24. Chen, C.-C. et al. Contributions of cation–π interactions to the collagen triple helix stability. *Arch. Biochem. Biophys.* **508**, 46–53 (2011).
25. Xu, F., Zahid, S., Silva, T. & Nanda, V. Computational design of a collagen A:B:C-type heterotrimer. *J. Am. Chem. Soc.* **133**, 15260–15263 (2011).
26. Parmar, A. S., Joshi, M., Nosker, P. L., Hasan, N. F. & Nanda, V. Control of collagen stability and heterotrimer specificity through repulsive electrostatic interactions. *Biomolecules* **3**, 986–996 (2013).
27. Acevedo-Jake, A. M., Ngo, D. H. & Hartgerink, J. D. Control of collagen triple helix stability by phosphorylation. *Biomacromolecules* **18**, 1157–1161 (2017).
28. Zheng, H., Liu, H., Hu, J. & Xu, F. Using a collagen heterotrimer to screen for cation-π interactions to stabilize triple helices. *Chem. Phys. Lett.* **715**, 77–83 (2019).
29. Gauba, V. & Hartgerink, J. D. Surprisingly high stability of collagen ABC heterotrimer: evaluation of side chain charge pairs. *J. Am. Chem. Soc.* **129**, 15034–15041 (2007).
30. Zheng, H. et al. How electrostatic networks modulate specificity and stability of collagen. *Proc. Natl Acad. Sci. USA* **115**, 6207–6212 (2018).
31. Fallas, J. A. & Hartgerink, J. D. Computational design of self-assembling register-specific collagen heterotrimers. *Nat. Commun.* **3**, 1087 (2012).
32. Acevedo-Jake, A. M., Clements, K. A. & Hartgerink, J. D. Synthetic, register-specific, AAB heterotrimers to investigate single point glycine mutations in Osteogenesis imperfecta. *Biomacromolecules* **17**, 914–921 (2016).
33. Clements, K. A., Acevedo-Jake, A. M., Walker, D. R. & Hartgerink, J. D. Glycine substitutions in collagen heterotrimers alter triple helical assembly. *Biomacromolecules* **18**, 617–624 (2017).
34. Giddu, S., Xu, F. & Nanda, V. Sequence recombination improves target specificity in a redesigned collagen peptide abc-type heterotrimer. *Proteins* **81**, 386–393 (2013).
35. Parmar, A. S. et al. Design of net-charged abc-type collagen heterotrimers. *J. Struct. Biol.* **185**, 163–167 (2014).
36. Egli, J., Erdmann, R. S., Schmidt, P. J. & Wennemers, H. Effect of N- and C-terminal functional groups on the stability of collagen triple helices. *Chem. Commun.* **53**, 11036–11039 (2017).
37. Keshwani, N., Banerjee, S., Brodsky, B. & Makhatadze, G. I. The role of cross-chain ionic interactions for the stability of collagen model peptides. *Biophys. J.* **105**, 1681–1688 (2013).
38. Acevedo-Jake, A. M., Jalan, A. A. & Hartgerink, J. D. Comparative NMR analysis of collagen triple helix organization from N- to C-termini. *Biomacromolecules* **16**, 145–155 (2015).
39. O'Leary, L. R. *Multi-Hierarchical Self-Assembly of collagen mimetic peptides into AAB type heterotrimers, nanofibers and hydrogels driven by charged pair interactions*. PhD thesis, Rice University (2011).
40. Deprez, P., Doss-Pepe, E., Brodsky, B. & Inestrosa, N. C. Interaction of the collagen-like tail of asymmetric acetylcholinesterase with heparin depends on triple-helical conformation, sequence and stability. *Biochem. J.* **350**, 283–290 (2000).
41. Shah, N. K., Sharma, M., Kirkpatrick, A., Ramshaw, J. A. M. & Brodsky, B. Gly-Gly-containing triplets of low stability adjacent to a type III collagen epitope. *Biochemistry* **36**, 5878–5883 (1997).

Methods

Peptide synthesis. Peptides were synthesized using standard Fmoc-protected amino acids with a low-loading Rink amide MBHA resin to give C-terminal amidation. A mixture of 25% v/v piperidine in dimethylformamide (DMF) was used for deprotecting steps. Coupling was performed using 2-(1H-7-azabenzotriazol-1-yl)-1,1,3,3-tetramethyl uranium hexafluorophosphate methanaminium (HATU) and diisopropylethylamine (DiEA) in DMF in the ratio 1:4:4:6 (resin:amino acids:HATU:DiEA). Acetylation of the N terminus was performed twice with an excess of acetic anhydride with DiEA in dichloromethane (DCM). Cleavage was performed with 10% v/v scavengers (anisole, triisopropylsilane, H_2O and ethanedithiol) in trifluoroacetic acid (TFA). For the ABC system, the peptides were synthesized with ^{15}N isotopically enriched glycine at position 21 to enable NMR experiments.

Peptide purification. TFA was removed from the reaction mixture by evaporation under nitrogen. Cold diethyl ether was used to triturate the crude peptide. After centrifugation, the crude pellet was washed with cold diethyl ether twice. Peptides were dissolved in H_2O to a concentration of 22 mg per ml and filtered before purification by reverse phase high-pressure liquid chromatography (HPLC) with water-acetonitrile with 0.05% TFA at a gradient of 0.7% per min on a 19 mm × 250 mm C-18 column. Samples were roto-evaporated to remove acetonitrile and were then lyophilized. Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) was used to confirm peptide mass. Pure samples were analysed at 1 mM concentration with a 2.1 mm × 50 mm C-18 column at a gradient of 14% per min water-acetonitrile with 0.05% TFA.

Sample preparation. Peptide samples were dissolved in MilliQ water with concentration determined by mass. They were diluted to 3 mM peptide in 10 mM phosphate buffer with pH 7.0, and were then preheated at 85 °C for 15 min. The samples were cooled to room temperature (20–25 °C) and were then stored overnight at 5 °C. They were diluted to 0.3 mM with water just before performing circular dichroism spectroscopy.

Circular dichroism spectroscopy. Circular dichroism spectroscopy was performed on a spectropolarimeter equipped with temperature control. 200 µl of the 0.3 mM peptide sample were transferred to a quartz cuvette with a pathlength of 0.1 cm. Wavelength scans were performed between 200 and 250 nm. The maximum near 225 nm was then followed as the temperature increased from 5 °C to 85 °C at 10 °C per h. Using the Savitzky–Golay smoothing algorithm, the first derivative curve was calculated. The minimum of the first derivative was defined as the melting temperature (T_m). The molar residue ellipticity (MRE) was calculated as previously reported[17].

Interpreting circular dichroism experiments in the context of collagen mimetic peptides. When analysing collagen mimetic peptides, circular dichroism experiments can be used to determine the composition of a combination of peptides that fold into a triple helix. Seven samples must be analysed for a ternary combination of peptides (A, B and C) to be thoroughly characterized: the three unary samples (A, B and C), the three binary mixtures thereof and the ternary mixture. Each level of peptide combinations build on the information gleaned from the previous level. Analysis of the unary samples provides information about the stability of the homotrimeric triple helices. New transitions viewed in the binary samples indicate the presence of binary composed heterotrimers, and transitions unique to the ternary mixture indicate a ternary composed heterotrimer. Besides only providing information about triple helical composition, sometimes there is ambiguity even in the helical composition due to overlapping similar transitions between different combinations of peptides. It is for these reasons that NMR can be a particularly useful tool for the elucidation of triple helical structure.

Axial–lateral interaction deconvolution. Peptides of $(POG)_8$, $(POG)_4(XOG)(POG)_3$, $(POG)_3(PYG)(POG)_4$, $(POG)_4(XYG)(POG)_3$ and $(POG)_3(PYGXOG)(POG)_3$ sequences were synthesized and characterized for each pair of interactions studied, unless otherwise noted. For this analysis, we did not use previously published melting temperatures in order to ensure consistency in our calculations. Homotrimers contain three units of each residue present in the peptide. The XYG homotrimer contains two lateral interactions, and the YGX homotrimer contains two axial interactions and one lateral interaction. The melting temperatures for each homotrimer can be represented by equations (1) and (2). Rearranging these equations allows us to solve for the value of the specific lateral or axial interaction given by equations (3) and (4).

$$Tm_{(XYG)} = Tm_{(POG)} + 3\times\Delta Tm_{(Xaa)} + 3\times\Delta Tm_{(Yaa)} + 2\times lat_{(YX)} \quad (1)$$

$$Tm_{(YGX)} = Tm_{(POG)} + 3\times\Delta Tm_{(Xaa)} + 3\times\Delta Tm_{(Yaa)} + lat_{(YX)} + 2\times ax_{(YX)} \quad (2)$$

$$lat_{(YX)} = (Tm_{(POG)} + 3\times\Delta Tm_{(Xaa)} + 3\times\Delta Tm_{(Yaa)} - Tm_{(XYG)})/(-2) \quad (3)$$

$$ax_{(YX)} = (Tm_{(POG)} + 3\times\Delta Tm_{(Xaa)} + 3\times\Delta Tm_{Yaa} - Tm_{(YGX)} - lat_{(YX)})/(-2) \quad (4)$$

Algorithmic design. SCEPTTr was written in Java using the Eclipse development environment. Scoring parameters were initially optimized by a genetic algorithm for the minimum value of $\Sigma(T_{m,measured} - T_{m,predicted})^2$. Details are provided in the Supplementary Discussion (see the section 'Optimization of SCEPTTr').

NMR. For the sample preparation, 450 µl of the 3 mM peptide samples were mixed with 50 µl D_2O and less than 0.5 mg of trimethylsilylpropionic acid as a reference.

Two-dimensional (2D) ^1H–^{15}N HSQC and three-dimensional (3D) ^1H–^1H–^{15}N NOESY HSQC experiments were performed on an 800 MHz Bruker spectrometer containing a cryogenic probe. The nitrogen carrier frequency was set at 108 ppm for the 3D experiment and at 112 ppm for the 2D experiments. The proton carrier frequency was set to match that of the water signal. All experiments were performed at 25 °C. The ^1H–^{15}N HSQC experiments were collected using 1,922 × 256 complex points for each of 8 scans with 32 preceding dummy scans. A sweep width of 12 ppm was used in the ^1H dimension and 25 ppm in the ^{15}N dimension. The 3D ^1H–^1H–^{15}N NOESY HSQC experiment was collected with a 90 ms mixing time, 4 scans and 2,048 × 32 × 256 complex points with 64 dummy scans. The sweep width was 12 ppm in the direct dimension, 10 ppm in the ^{15}N dimension and 8 ppm in the indirect dimension. Raw NMR data were processed using TopSpin 4.0 software. A baseline correction (qpol for 2D experiments and sfil for the 3D experiment) and a window multiplication function were applied (sine squared function for the direct dimension of the 2D experiments and all three dimensions of the 3D experiment, and sine function for the indirect dimension of the 2D experiments). Data were zero-filled to process 2,048 × 1,024 complex points in the HSQC experiments and 2,018 × 128 × 1,024 complex points in the NOESY HSQC experiment. Each dimension for each experiment was Fourier-transformed and phase-corrected.

Interpreting ^1H–^{15}N HSQC NMR experiments in the context of collagen mimetic peptides. HSQC NMR experiments are used to determine the relationship between two atoms bound to each other and their respective NMR shifts. In these experiments, a peak is indicative of a pair of atoms bonded together, with the Cartesian coordinates indicating the NMR shifts of the two atoms involved. When analysing glycine-labelled collagen mimetic peptides, ^1H–^{15}N HSQC experiments must be run in a similar manner as circular dichroism spectroscopy. In other words, each peptide must be analysed alone and mixed with the other peptides to understand and study the results of the ternary mixture. Each labelled glycine in a unique chemical environment is expected to result in a unique peak in the HSQC spectrum. Examination of the unary peptide samples enables differentiation of the separate monomer peaks. Owing to the equilibrium between a monomer and a trimer in any solution of CMPs, it is expected that these monomer peaks will be evident in every mixture containing the given peptide. Further, monomer peaks were confirmed by identifying peaks that were taller and broader than peaks arising from a triple helical sample due to tumbling rates skewing towards a monomer, equilibrium rates skewing towards a monomer (particularly in many unary samples) and greater solvent exposure for these amides. When considering triple helical peaks, again, each unique chemical environment is expected to result in a unique peak; therefore, each triple helix is expected to be represented by three HSQC peaks (in systems in which each peptide contains one labelled glycine). However, these peaks may not always be distinguishable if there is overlap between the peaks themselves or an overlap with other peaks in the sample, including monomer peaks. After running the experiments for all seven combinations of peptides, a comparison of the spectra reveals (as with circular dichroism spectroscopy) the composition of the triple helix observed in each mixture. However, despite the limited ability of circular dichroism spectroscopy to detect mixtures of triple helices, an HSQC spectrum displaying three monomeric peaks and three triple helical peaks can be safely designated as being that of a pure species exhibiting a single composition and single register.

Interpreting 3D ^1H–^1H–^{15}N NOESY HSQC NMR experiments in the context of collagen mimetic peptides. When analysing glycine-labelled collagen mimetic peptides, 3D ^1H–^1H–^{15}N NOESY HSQC experiments are highly beneficial in determining the registration of a triple helix. This experiment produces data for intensities plotted inside a 3D cube of NMR shifts. Viewing this cube from the HSQC face appears identical to the HSQC experiment and viewing it from the NOESY face appears identical to the ^{15}N-filtered NOESY. NOESY provides information about ^1H-^1H interactions through space. A cross-peak between two hydrogens can be expected if they are within ~5 Å of each other. Filtering the NOESY with the ^{15}N label simplifies the experiment to show only the hydrogens that are bound to ^{15}N, which makes the analysis easier. The 3D experiment is more powerful than the simple filtered experiment because it allows separation of the NOESY by creating slices perpendicular to the HSQC plane. In other words, the 3D experiment makes it possible to select a particular ^{15}N shift in the HSQC spectrum and analyse the NOESY at only that ^{15}N signal. This enables the separation and analysis of two glycines that have similar ^1H shifts but different ^{15}N shifts. For collagen helices, labelled glycines have three main NOESY peaks that can be used to determine the registration of the triple helix: (1) NOEs to other nearby glycine amide hydrogens; (2) NOEs to the Cα hydrogens of the sequentially preceding Yaa position amino acid; (3) and NOEs to the Cα hydrogens

of Yaa-position amino acids on different strands. In (1), each glycine amide proton will possess an NOE to the amide of the glycines in each cross-section of the triple helix adjacent to the glycine of interest. With three labelled glycines, this information can be useful for identifying the middle strand; however, the leading and trailing strands cannot be differentiated using this analysis. Additionally, at times, these amide peaks will overlap with one another, resulting in NOEs that are buried under the main peaks and therefore unhelpful. Using expected peaks (2) and (3) for each glycine can often help to rectify this resulting ambiguity. Peak (2) is always a large peak in the region of ~4–5 ppm because the Cα hydrogen of the Yaa-position amino acid preceding a glycine is ~2.2 Å from the amide proton. Peak (3), which corresponds to the Yaa-position amino acid in the same helical cross-section as the glycine of interest, is much smaller because this NOE is over a larger distance, ~3.2 Å. The register is evaluated using these peaks, based on the knowledge that peak (3) of the leading strand corresponds to the same Cα hydrogen as peak (2) of the middle strand, and that peak (3) of the middle strand corresponds to the same Cα hydrogen as peak (3) of the trailing strand. Thus, the order of the strands can be determined from the Cα hydrogen NOE record.

Data availability

The data that support the findings of this study are available within the paper, in the Supplementary Information and from the corresponding author. In particular, the NMR raw data are available on request due to the size of the files and the lack of appropriate public repositories for raw multi-dimensional NMR data.

Code availability

A compiled standalone Java application is available for download as Supplementary Software 1. The genetic algorithm used to optimize SCEPTTr requires human input to further optimize the values used. For this reason, it is available from the corresponding author upon reasonable request.

Acknowledgements

This work was funded by the National Science Foundation (CHE 1709631) and the Robert A. Welch Foundation (C1557). I.-C.L. was supported by the Stauffer-Rothrock Fellowship. We thank the NMR and Drug Metabolism Core at Baylor College of Medicine for access to the 800 MHz NMR spectrometer and K. R. MacKenzie for assistance in acquiring NMR spectra.

Author contributions

D.R.W., S.A.H.H., I.-C.L. and K.J.G. designed, synthesized and characterized deconvolution peptides. D.R.W. assembled the peptide library, designed, wrote and optimized SCEPTTr, performed and analysed NMR experiments, and co-wrote the manuscript. C.M.P. designed and synthesized the novel ABC system and characterized it using circular dichroism spectroscopy. J.D.H. supervised the research, evaluated all of the data and co-wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41557-020-00626-6.

Correspondence and requests for materials should be addressed to J.D.H.

Peer review information *Nature Chemistry* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.D.H.