nature communications

Article

https://doi.org/10.1038/s41467-024-54046-y

Deciphering the folding code of collagens

Received: 10 February 2024

Accepted: 30 October 2024

Published online: 19 March 2025

Check for updates

Jean-Daniel Malcor[1], Noelia Ferruz[2,4,6], Sergio Romero-Romero[2,5,6], Surbhi Dhingra[2], Vamika Sagar[3] & Abhishek A. Jalan[2,3]

Collagen proteins contain a characteristic structural motif called a triple helix. During the self-assembly of this motif, three polypeptides form a folding nucleus at the C-termini and then propagate towards the N-termini like a zip-chain. While polypeptides from human collagens contain up to a 1000 amino acids, those found in bacteria can contain up to 6000 amino acids. Additionally, the collagen polypeptides are also frequently interrupted by non-helical sequences that disrupt folding and reduce stability. Given the length of polypeptides and the disruptive interruptions, compensating mechanisms that stabilize against local unfolding during propagation and offset the entropic cost of folding are not fully understood. Here, we show that the information for the correct folding of collagen triple helices is encoded in their sequence as interchain electrostatic interactions, which likely act as molecular clamps that prevent local unfolding. In the case of humans, disrupting these electrostatic interactions is associated with severe to lethal diseases.

Collagens are highly abundant human proteins that provide structure and strength to tissues and bind cell-surface and secreted proteins to regulate key biological processes, including tissue homeostasis and blood clotting. All collagen proteins contain a characteristic domain called a triple helix composed of three supercoiled polypeptides. In humans, collagen genes encode 44 polypeptides, also called α-chains, that self-assemble into 28 distinct collagen types (I-XXVIII)[1]. The α-chains contain a repetitive three amino acid sequence Gly-Xaa-Yaa, where glycine occupies every third position and Xaa and Yaa are more frequently non-glycine amino acids. Upon trimerization, the recurrent glycines sequester into a tightly packed core of the triple helix[2]. Consequently, their mutation to other amino acids perturbs the structure resulting in delayed folding and reduced stability[3,4]. This is the leading cause of heritable collagen-related diseases[5]. Direct study of collagen folding and stability is complicated by the propensity of native collagens to form insoluble aggregates in vitro. Fortunately, synthetic peptides containing sufficient Gly-Xaa-Yaa repeats intrinsically self-assemble into a triple helix. Investigations using such peptides suggest that triple helices fold via a nucleation-zipper mechanism, where three peptides nucleate at the C-terminal and then propagate towards the N-terminal like a zip-chain[6].

The α-chains of human collagens contain up to 330 Gly-Xaa-Yaa repeats. Intuitively, until the propagating triple helix has reached a critical length sufficient to sustain further folding, the propagation phase is expected to be entropically disrupted due to the substantial length of the α-chains. In some collagens, the perfect triplet repeat pattern is also frequently interrupted by non-helical sequences that lower triple helix thermal stability and disrupt folding[7]. In such cases, the triple helix must renucleate after the interruption while also compensating for the loss in local stability and folding rate. A collagen-specific chaperone heat shock protein (Hsp) 47 resident in the endoplasmic reticulum has been shown to recognize unique sites on the triple-helical domain of some human collagens[8–10]. This interaction between Hsp47 and collagen triple helices is believed to provide stability against local unfolding while offsetting the entropic loss[11]. However, several experimentally observed features of collagen folding are not fully consistent with the Hsp47-chaperoned folding paradigm. Prominent among these is the observation that Hsp47 does not recognize some types of collagens, or does so only weakly[12]. Additionally, on a structural level, polypeptides within a triple helix adopt a one-amino acid staggered alignment with respect to each other. But, non-canonical polypeptide alignments where the chains are offset by more than one amino acid are not uncommon[13]. Hsp47 binding does not explain how polypeptides avoid incorrect alignment during folding. Hsp47 binding also does not explain how triple helices in native

[1]Laboratory of Tissue Biology and Therapeutic Engineering, CNRS UMR 5305 University of Lyon, Lyon, France. [2]Department of Biochemistry, University of Bayreuth, Bayreuth, Germany. [3]Department of Biomaterials, University of Bayreuth, Bayreuth, Germany. [4]Present address: Centre for Genomic Regulation, Barcelona, Spain. [5]Present address: Department of Biochemistry and Structural Biology. Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, Mexico City, Mexico. [6]These authors contributed equally: Noelia Ferruz, Sergio Romero-Romero. e-mail: jalan@cantab.net

collagen re-nucleate after an interruption and compensate for the loss in stability and folding rate. Moreover, in addition to eukaryotic collagens, eukaryotes[14], prokaryotes[15], and viruses[16] also encode collagen-like proteins containing the characteristic Gly-Xaa-Yaa repeats that form stable and functional triple helices[17–19]. While human α-chains are up to 330 triplets long, collagen-like polypeptides in prokaryotes can reach lengths of up to 2000 triplets and contain a greater number of interruptions[15]. How such long polypeptides mitigate local unfolding during the propagation phase while also compensating for the disrupting effect of interruptions is currently not understood. The information for the correct folding of most proteins is encoded in their amino acid sequence. Our results here suggest this to be true for collagens as well.

Here we show that human collagens and collagen-like proteins in archaea, bacteria, and eukarya (excluding human collagen orthologues), and viruses contain amino acid motifs capable of forming geometrically specific lysine-glutamate and lysine-aspartate salt bridges. The salt bridges decrease the unfolding rate i.e. increase kinetic stability of model triple-helical peptides as well as several native collagens tested here. We find that the 28 known collagen subtypes contain an average of 50 salt bridges each. By comparison, only a few binding sites for Hsp47 have formally been identified in human type II and III collagens[20]. Importantly, incorrect alignment of the chains in the respective triple helices dramatically decreases the number of possible salt bridges in all collagen subtypes. We also find that salt bridges are distributed throughout the length of the triple-helical domain of human collagens but their density is especially higher in the vicinity of non-collagenous interruptions. This combined with their ability to increase kinetic stability suggests that salt bridges act as local electrostatic clamps that lock the folded regions of the triple helices in place and allow the remaining regions to propagate. Therefore, our results provide a general mechanism for how collagen triple helices avoid local unfolding while precluding unproductive folding intermediates. We also find that mutations that disrupt salt bridges are associated with severe or lethal diseases, which has important consequences for understanding why some mutations in human collagens cause more severe phenotypes than others.

Results

Triplets that form salt bridges are anomalously frequent in collagen and collagen-like proteins

Analysis of human fibrillar collagens, representing 9 of the 44 collagen α-chains, has previously revealed that several Yaa-Gly-Xaa triplets including KGE and KGD occur in the triple-helical domain with a frequency greater than that expected from a random distribution of amino acids[21]. Such an anomalous frequency suggests selection pressure and thus a prominent biological role in folding and/or function. These triplets possess an ammonium group (on the side chain of lysines) and a carboxylate group (on the side chains of aspartate or glutamate) in close proximity, making them likely to participate in electrostatic salt bridge interactions. We revisited the analysis of KGD or KGE triplet frequency within all 44 collagen α-chains in humans. Additionally, we also analyzed the triplet frequency in collagen-like proteins from archaea, bacteria, eukarya, and viruses to understand how these compare to the human collagens.

Considering the 20 natural amino acids, there are 400 theoretical combinations of Xaa and Yaa paired residues in the Yaa-Gly-Xaa triplets. While glycines present in the Xaa and Yaa positions have traditionally been considered part of the triple-helical domain, we here consider them as interruptions. The presence of glycines in the Xaa or Yaa position disrupts the tightly wound helical structure at the site of interruption. This is evident from the fact that the backbone dihedral angles of this glycine and residues surrounding it deviate away from the canonical polyproline type II area of the Ramachandran plot[2,22]. This disruption causes a loss of the canonical pattern of hydrogen-bonds between the chains, which reduces the interchain van der Waals contact area and dramatically reduces the thermal stability of the triple helices[23]. Thus, the Gly-Yaa-Gly-Gly and Gly-Gly-Xaa-Gly motifs, abbreviate henceforth as GYGG and GGXGG, should more aptly be considered interruptions instead of a part of the triple-helical domain.

Excluding glycine, there are 361 residue pair combinations possible in a Yaa-Gly-Xaa triplet. Considering that 6 or more contiguous repeats of Yaa-Gly-Xaa triplets have previously been defined as a triple-helical domain[24], we identified 468 collagen domains in humans while archaea, bacteria, eukarya, and viruses accounted for 58002 collagen-like protein domains (Supplementary Table 1 and Supplementary Data 1). In order to understand the frequency of triplets in the collagen and collagen-like domains, the occurrence of each Yaa-Gly-Xaa triplets was first predicted based on the observed individual frequency of amino acid residues in the Xaa and Yaa positions within the sequences of human collagen as reported previously[21] (see methods for details). This prediction was then compared to the observed occurrence of Xaa and Yaa pair combinations in Yaa-Gly-Xaa triplets.

Triplets with Z-scores greater than 3 or less than −3 (more than three standard deviations away from the mean of the difference between observed and predicted values) were considered as anomalously frequent. In order to understand if triplets with anomalously high frequency also have high relative abundance, Z-scores were plotted against relative abundance, which is defined as the percent observed instance of a triplet against the total number of triplets in collagens or collagen-like proteins. As shown in Fig. 1A, KGE and/or KGD triplets are Z-score outliers and thus anomalously frequent not only in the α-chains of human collagens but also in collagen-like proteins in archaea, eukarya, bacteria, and viruses. The KGE and KGD triplets also have high relative abundance ranging between 3 and 6% in both collagens and collagen-like proteins. In the case of viral collagen-like proteins, KGD accounts for a stupendous 13% of all observed triplets. In eukaryotes, the PGP triplet, or OGP following the post-translational modification of proline into 4(*R*)-hydroxyproline (noted O), is by far the most abundant. It should be noted that the post-translational modification of proline in Yaa position is presumed for many collagen types rather than experimentally demonstrated. In any case, a succession of OGP triplets assembles into the most thermally stable triple helix, with any individual mutation at the Xaa or Yaa position resulting in destabilization. In particular, in model triple-helical peptides, a hydroxyproline to lysine mutation decreases the triple helix thermal stability by 10 °C, and a proline to glutamate or aspartate mutation destabilizes it by 7 °C or 4 °C respectively[25]. Yet, mutating both hydroxyproline and proline to lysine and glutamate results in a moderate thermal stability decrease of 5–8 °C, suggesting a compensating stabilization mechanism stemming from salt bridge formation[26].

Threonine-containing triplets TGA and TGP are also found to be anomalously frequent with high relative abundance in archaea, bacteria, and viruses but not in human collagen or collagen-like proteins in eukarya. Bacteria lack the enzyme prolyl hydroxylase required for post-translational modification of proline to hydroxyproline. The high abundance of threonine in place of hydroxyproline has been rationalized based on the ability of bacteria to glycosylate threonine[27] which presumably increases triple-helical stability via water-mediated hydrogen bonds[28].

We next analyzed the frequency of Yaa-Gly-Xaa triplets that surround interruptions in human collagens and collagen-like proteins, following a protocol developed by Bella et al.[24]. In this nomenclature, the interruptions are denoted GnG according to sequence length, where n is the length of the interruption up to 15 amino acids between two glycines. For example, a deletion of one amino acid from the triplet repeat sequence is denoted as G1G, and an insertion of one amino acid is noted as G3G. Importantly, we consider the presence of glycine in the Xaa or Yaa positions as interruptions. We abbreviate the

https://doi.org/10.1038/s41467-024-54046-y

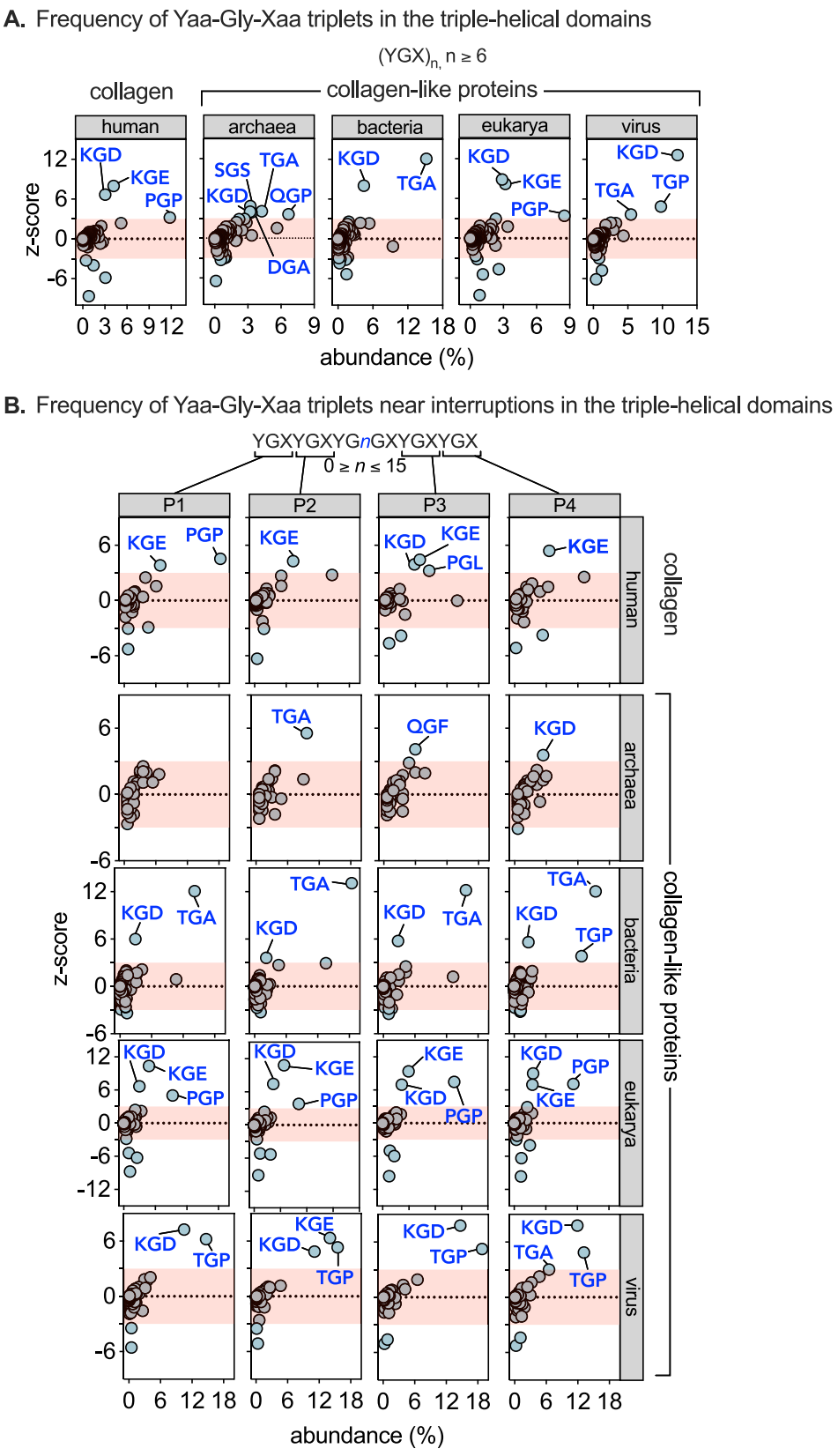

Fig. 1 | Electrostatically charged triplets are anomalously frequent in collagen and collagen-like proteins. Frequency of Yaa-Gly-Xaa triplets in the uninterrupted triple-helical domains (**A**) and surrounding the interruptions (**B**) in the 44 known human collagen α-chains and in the collagen-like proteins from humans, viruses and the three superkingdoms. In the case of interruptions G*n*G, where $0 \geq n \leq 15$, the triplet frequency was determined in the four triplets (labeled P1-P4) surrounding the interruptions.

resulting GXGG and GGYG motifs as G2G interruptions. Fibrillar collagens (especially collagen type III) contain many instances of such interruptions.

Including G2G, we identified 393 interruptions in human collagens and 30124 interruptions in collagen-like proteins after removing duplicates (Supplementary Table 1 and Supplementary Data 2). In human collagens, one duplicate interruption is found in both type II and type V collagens, making a total of 394 interruptions. Next, we determined the frequency of Yaa-Gly-Xaa triplets focusing only on the sequence surrounding the interruptions. Starting from the N-termini, the two triplets on either side of the interruptions were labeled P1-P4. As shown in Fig. 1B, KGE and/or KGD are anomalously frequent in P1-P4 positions surrounding interruptions in not only human collagens but also collagen-like proteins from bacteria, eukarya, and viruses. Intriguingly, we find OGP triplet to be anomalously frequent in the triple-helical domains in general but it does not appear with anomalous frequency near the interruptions in humans. Archaeal collagen-like proteins show anomalous frequency of KGD in only the P4 position. The relative abundance of the KGE and KGD triplets surrounding interruptions in collagen-like proteins of bacteria and eukarya ranges between 2 and 6%. Among all analyzed groups, viruses show the highest relative abundance of these triplets at 10-12% in all four positions surrounding the interruptions.

The consistent observation that KGE and KGD triplets are anomalously frequent in the triple-helical domains of human collagens and collagen-like proteins in the four groups and that they are also enriched in the triplets surrounding interruption sites (in P1 to P4 for bacteria, eukarya and virus, and in P4 for archaea) suggests a prominent and likely common role in stability, folding and/or function.

KGE and KGD triplet-mediated salt bridges are widely distributed in all human collagens

We established the anomalously high frequency of KGE and KGD triplets in human collagens via analysis of single collagen α-chains. However, which of these triplets are capable of forming salt bridges upon trimerization, how many such salt bridges are feasible in a given collagen subtype, what is their distribution within the triple-helical domain and how they are located vis-à-vis interruption is currently not known. To this end, we aligned the 44 collagen α-chains into the triple-helical stoichiometries observed in humans[1] and determined the number and location of salt bridges in the triple-helical domains and also surrounding the interruptions (Supplementary Fig. 1). It should be noted that we aligned human collagens using a consideration of previously published sequence-based information and experimental analysis of the binding of collagen to other proteins (see methods and Supplementary Table 2). In cases where such considerations were not available, we aligned sequences by optimizing the length of the triple-helical domain. Given that direct experimental verification of such alignments is currently not available, we refer to ours as proposed alignments.

The lysine and aspartate/glutamate residues capable of forming a salt bridge in a triple helix are constrained by sequence. In order to understand these constraints, we refer to the three staggered chains of the triple-helix as leading (staggered towards the N-terminal end), middle, or trailing (staggered towards the C-terminal end). In this scheme, the i^{th} lysine of the leading and middle chains form salt bridges with $i+2$ aspartate or glutamate of the middle and trailing chain, respectively (Fig. 2A). The remaining lysine on the trailing chain can also form a salt bridge with the aspartate or glutamate occupying the $i+5$ positions in the leading chain. Salt bridges that conform to the three sequence constraints are identical with respect to the spatial condition for interaction. However, the $i \rightarrow i+5$ salt bridge is formed by residues that may or may not be present within a KGE or KGD triplet. A search of the aligned triple-helical domains of human collagens shown in Supplementary Fig. 1 for pairs of lysine and aspartate/glutamate residues that satisfy any of the three sequence constraints revealed 1553 salt bridges (Fig. 2B and Supplementary Table 3). Of these, 21% are $i \rightarrow i+5$ salt bridges that do not originate from the KGE or KGD triplets. This emphasizes the need to search aligned triple-helical sequences rather than relying solely on instances of KGE and KGD triplets to identify potential salt bridge interactions.

The observed 1553 salt bridges are distributed across the 28 human collagens with an average of ~50 in each subtype. A visual inspection of their distribution shown in Fig. 2C suggests that they are present throughout the length of the triple-helical domain. In order to further understand this distribution, we divided the triple-helical domains of aligned collagen sequences into four equal parts and counted the number of salt bridges in each quarter. As shown in Supplementary Fig. 2, the C-terminal quarter accounts for ~50% (757 out of 1553) of all salt bridges. This highly skewed distribution of salt bridges assumes significance in view of the C-terminal nucleation-propagation model currently accepted for the folding of collagen triple helices.

Misalignment of α-chains reduces the number of possible salt bridges

Polypeptides within a triple helix are staggered by one amino acid with respect to each other. This canonical alignment maximizes interchain hydrogen bonds and allows glycine residues to sequester into the core of the triple helix[2]. In theory, the polypeptides can be staggered by an arithmetic series of 1, 4, 7, 10,… amino acids while retaining the tightly-packed triple-helical structure. However, non-canonical alignments of more than 1 amino acid would result in loss of hydrogen bonds at the termini. Thus, triple helices containing non-canonically aligned polypeptides have not been experimentally observed in short collagen peptides. However, given that the triple-helical domains of human collagens contain up to 1000 amino acids and >900 hydrogen bonds, the free energy difference between the canonical and non-canonically aligned triple helices is expected to be small. Thus, if only hydrogen bond and Van der Waals packing are considered, the canonical and non-canonical staggers of collagens would be separated by low energy barriers, resulting in frequent misfolding via misalignment of chains. It remains to be understood how collagens avoid such local folding traps and find the global minimum.

As noted in the previous section, all human collagens contain an average of 50 salt bridges in each collagen subtype. We find that the number of possible salt bridges decreases dramatically upon intentionally misaligning either the middle or the trailing chain by just 4 residues (Fig. 3 and Supplementary Table 3). The average loss of salt bridges upon misaligning the middle and trailing chain is 40 and 30%, respectively. As a representative example, the $\alpha3\alpha4\alpha5$ heterotrimer of collagen type IV loses 8% of salt bridges, while homotrimeric collagen type X loses 94% of salt bridges upon misalignment of the middle chain by 4 residues. The loss in salt bridges is expected to increase the free energy gap between the native and the competing misfolded states, thus ensuring that only the state with the canonical alignment is populated. Importantly, given that 50% of all salt bridges are concentrated in the C-terminal quarter of the triple-helical domains, the decision for correct alignment of the polypeptides is likely made early during the propagation phase.

Collagen subtypes with more interruptions also contain more salt bridges

We find that the number of salt bridges in each collagen subtype is positively correlated to both the number of interruptions and the length of the triple helical domain (Supplementary Fig. 3a–c). In physical terms, collagens with more interruptions or longer triple helical domains also contain a greater number of salt bridges. As a representative example, the $(\alpha1)_2\alpha2$ heterotrimer of collagen type I with 4 interruptions contains 35 salt bridges but that of collagen type IV with

https://doi.org/10.1038/s41467-024-54046-y

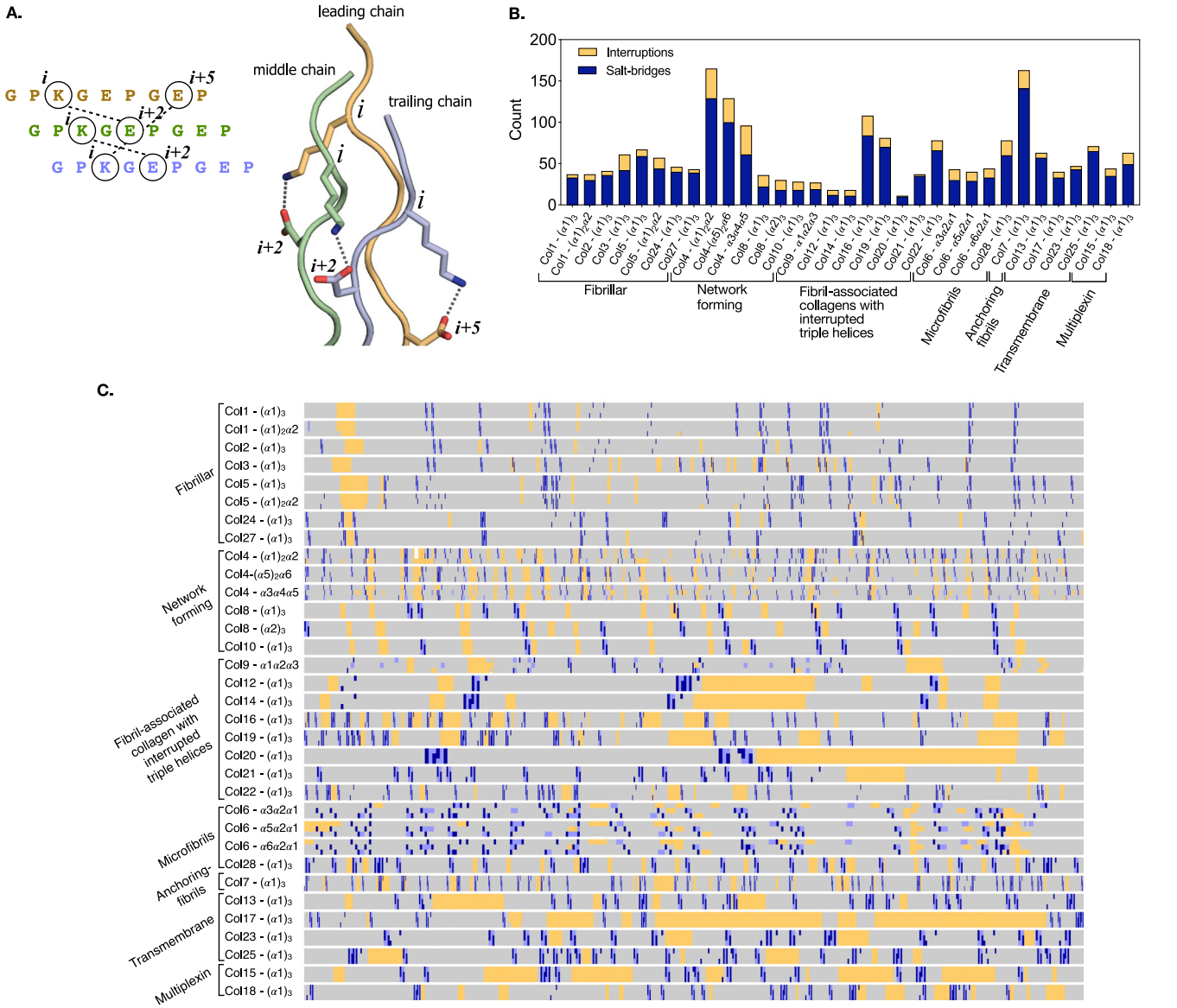

Fig. 2 | Salt bridges are abundant across all 28 human collagens. A Schematic depiction of the sequence constraint for salt bridge formation between the leading (orange), middle (green) and trailing (blue) chains of a triple helix [pdb: 6q3p[45]]. **B** The total number of salt bridges (blue) and interruptions (yellow) observed in the triple-helical domain of each collagen subtype. **C** A visual representation of the distribution of salt bridges (blue), KGE or KGD triplets (light blue), and interruptions (yellow) in the triple-helical domains of human collagens.

23 interruptions contains 129 salt bridges. Similarly, collagen type XX with the shortest triple helical domain (141 residues) among all collagens contains only 10 salt bridges while collagen type VII with the longest triple-helical domain (1380 residues) contains a staggering 141 salt bridges. While the number of salt bridges is strongly correlated to both the number of interruptions (Pearson correlation = 0.69, $P_{two\text{-}tailed} < 0.0001$, $\alpha = 0.05$) and triplets (Pearson correlation = 0.68, $P_{two\text{-}tailed} < 0.0001$, $\alpha = 0.05$), a weaker correlation is observed between the number of triplets and interruptions (Pearson correlation = 0.53, $P_{two\text{-}tailed} = 0.0014$, $\alpha = 0.05$). Thus, longer triple helices do not necessarily contain more interruptions.

By comparison, as shown in Supplementary Fig. 3D, E, the number of OGP triplets is strongly correlated to the length of the collagen triple helix (Pearson correlation = 0.78, $P_{two\text{-}tailed} < 0.0001$, $\alpha = 0.05$), but only weakly correlated to the number of interruption (Pearson correlation = 0.44, $P_{two\text{-}tailed} < 0.0001$, $\alpha = 0.05$). Interruptions disrupt the triple-helical structure[29] causing delayed folding and decreased overall stability[30]. Similarly, collagens with longer triple helical domains are expected to experience greater entropic disruption of folding during the propagation phase. We hypothesize that the increased abundance of salt bridges, but not OGP triplets, likely compensates for the disruptive effects of interruptions while both salt bridges and OGP triplets stabilize longer triple-helical domains.

Human collagen interruptions are flanked by salt-bridge "knots"

Of the 394 interruptions identified in the aligned triple-helical domains of human collagens, 58% contain salt bridges on the N- or C-termini or both (Supplementary Table 3). Close inspection suggests that the salt bridges seldom appear alone. We find that analogous to the cysteine knots in collagens[31], where two or more pairs of cysteine residues form disulfide bridges covalently linking all three chains, two or more pairs of interacting lysine and aspartate/glutamate residues also link all three chains. We call these salt bridge knots (Supplementary Fig. 4). We distinguish salt bridge knots from complex salt bridges observed in many globular proteins. As originally defined by Musafia et al.[32], in a complex salt bridge, a cationic or anionic residue simultaneously

https://doi.org/10.1038/s41467-024-54046-y

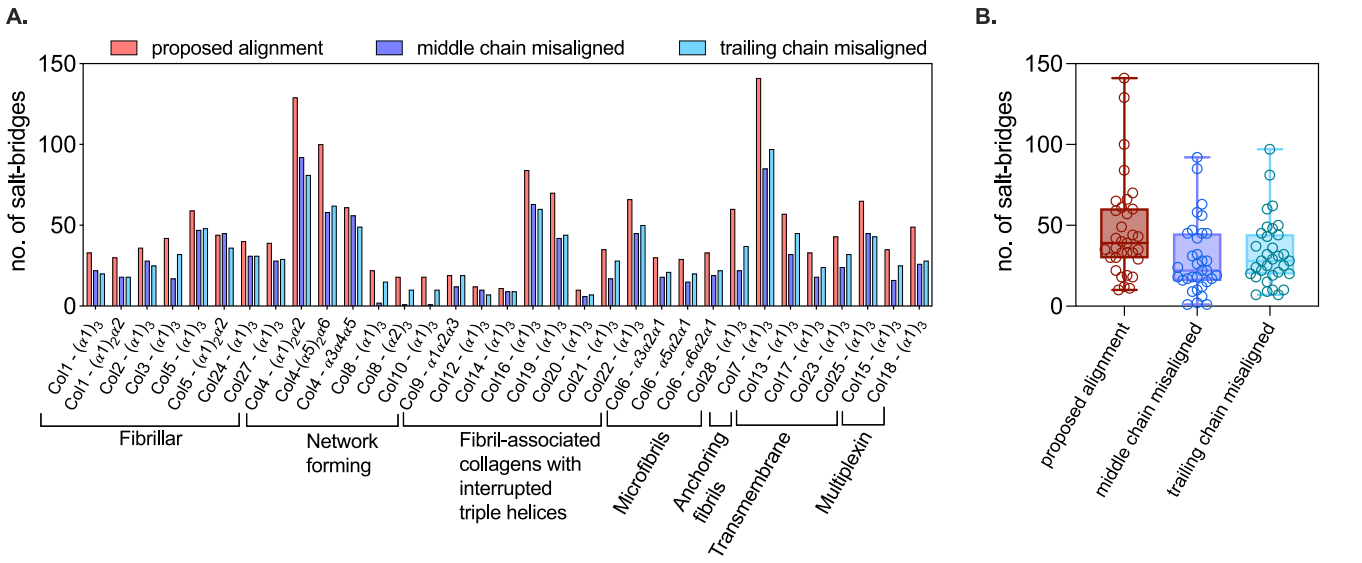

Fig. 3 | Misalignment of α chains reduces the number of salt bridges. Number of salt bridges observed in the proposed alignment shown in Supplementary Fig. 1 (red) and when the middle chain (purple) or trailing chain (cyan) are misaligned with a four-residue offset **A**. The box and whiskers plot in **B** show the aggregate loss in the number of salt bridges upon misalignment. The end limits of the error bars show the minimum and maximum number of salt-bridges, the box represents the lower and upper quartile of the total number of salt bridges and the horizontal line in the box denotes the median value.

interacts with two or more charged residues. In contrast, we define salt bridge knots as pairs of interacting cationic and anionic residues. As shown in Fig. 4, we identify knots containing up to eight salt bridges within the sequence space of four triplets.

A quantitative assessment shows that salt bridge knots are predominantly located close to interruptions. As noted in Supplementary Table 3, the proposed alignments of human collagens in this work reveals a total of 394 interruptions sites. The flanking P1-P4 regions of these interruptions account for 21% (394 * 4 triplets * 3 chains = 4728 triplets) of the 22811 triplets in the aligned collagen sequences but they contain 48% (228 out of 474) of all salt bridge knots. This suggests that more salt bridge knots are located close to interruptions than elsewhere in the triple helix. In addition, within the P1-P4 region, 19% of triplets are involved in salt bridge knots. This is significantly more than the percentage of triplets involved in salt bridge knots outside of the P1-P4 region (6%) and in all of the triple helical domains of human collagens (9%, one-way ANOVA, Tukey's post-test, Supplementary Fig. 5).

The general abundance of salt bridge knots in human collagens and in particular their preferential enrichment close to the interruptions is intriguing and suggests a prominent functional role. While we currently do not understand what this function might be, a survey of salt bridge abundance and stability in thermophilic, hyperthermophilic and halophilic organisms offers a clue. Proteins from these organisms frequently contain complex salt bridges, which cooperatively stabilize proteins i.e. the total increase in stability is more than that obtained from a simple sum of individual salt bridges[33,34]. Due to the extended rod-like topology and the sequence constraint for interaction, complex salt bridges are geometrically not feasible in triple-helices. It is plausible that salt bridge knots in collagens have evolved as an alternative with a role energetically similar to complex salt bridges in globular proteins.

KGE and KGD triplets form geometrically specific salt bridges

Salt bridges can be stabilizing or destabilizing[35,36], depending on the protein context. In a remarkable work, Kumar et al. have shown that the relative geometry of interacting cationic (ammonium or guanidinium) and anionic (carboxylate) head group determines whether or not a salt bridge is stabilizing[37]. This suggests that the stability conferred by a salt bridge is intimately linked to its geometry. Thus, we analyzed the geometry of salt bridges in the published crystal and solution structures of collagen triple helices using a parameterization previously developed for those in globular protein[38]. In this parameterization, the geometry of lysine-aspartate/glutamate salt bridges is defined by the angle between carboxylate oxygen and the Cε-Nζ atoms of lysine and the dihedral angle between the carboxylate oxygen and Cδ-Cε-Nζ atoms of lysine. The carboxylate group can adopt one of the three staggered configurations with respect to the tetrahedral ammonium group; gauche plus (*g*+), trans (*t*) and gauche minus (*g*–). Salt bridges in globular proteins are found to favor *g*+ or *g*- configurations. Additionally, the carboxylates of glutamate or aspartate can accept hydrogen bonds via either of the two nonbonded lone pairs of electrons called the syn and anti. Syn carboxylic acid is a weaker acid than anti and thus its conjugate base is more basic[39]. Despite the increased basicity, lysine-aspartate/glutamate salt bridges in globular proteins predominantly form hydrogen bonds via the anti-lone pair.

The stabilizing effect of both KGE/KGD salt bridges in a triple helix is well documented in the literature[21,40–43] and their molecular structure affirming the direct interaction between the lysine and aspartate/glutamate in model triple helical peptides has also been demonstrated by several independent research labs using crystallography[44,45] as well as NMR[46]. To the best of our knowledge, there is only one published molecular structure containing two lysine-glutamate salt bridges in collagen triple helices (pdb: 3t4f)[26]. Due to a lack of sufficient data points for lysine – glutamate salt bridges, we attempted to understand the geometric preference of salt bridges in collagens using published structures of triple-helical peptides containing lysine-aspartate salt bridges.

As shown in Supplementary Fig. 6 and Supplementary Data 3, the dihedral angle Cδ-Cε-Nζ---$O_{carboxylate}$ and angle Cε-Nζ---$O_{carboxylate}$ of salt bridges cluster around 180° and 90°, respectively. This suggests an overwhelming preference for the *trans* configuration, in contrast to the *g*+/*g*- preference in globular proteins. The geometric specificity also manifests in whether the syn or anti-lone pair interacts with the ammonium moiety. We parameterized this using the angle Oδ2-Oδ2---$Nζ_{lysine}$. Angles between 0 and 120° were classified as *syn* and those above 120° as *anti*. As shown in Supplementary Data 3, the majority of the interactions between the carboxylate and the ammonium group are via the more basic syn lone pair, again in contrast to globular

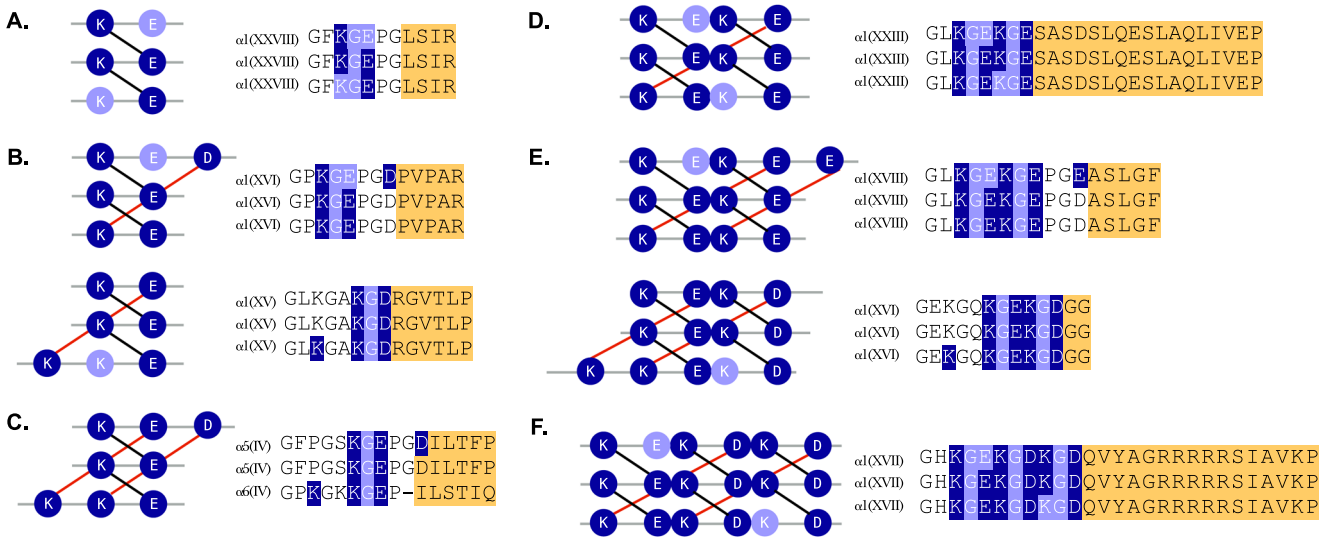

Fig. 4 | Salt bridge knots flank the interruptions sites in human collagens. Representative examples of salt bridge knots observed in human collagens containing two to six (**A**–**E**) and eight (**F**) salt bridges. Interacting pairs of lysine and aspartate/glutamate residues are shown in dark blue, the KGE and KGD triplets are shown in light purple and interruptions are shown in yellow. The interchain $i \rightarrow i+2$ salt bridges are shown with black line while the $i \rightarrow i+5$ salt bridges are shown in red. Although the representative examples shown in this figure contain interruptions only towards the C-termini, examples of salt bridge knots on the N-termini also abound (Supplementary Fig. 4).

proteins. Thus, our analysis suggests that KGD and KGE triplets also foster geometrically specific salt bridges but, in contrast to globular proteins, these salt bridges strongly favor trans configuration and nearly always interact via the more basic syn lone pair.

It should be noted that the direct experimental characterization of salt bridge geometry within native collagen molecules has been hindered due to the technical limitation of obtaining molecular structures of such complex, aggregation-prone and conformationally flexible proteins. We are aware of at least one instance where the molecular structure of collagen type I has been obtained via molecular dynamic simulations constrained with experimentally determined fiber diffraction parameters[47]. While its inspection suggests presence of several lysine-aspartate as well as lysine-glutamate salt bridges, we decided to exclude this structure from our analysis. This is primarily because the alignment of the α1 and α2 chains in the heterotrimer structure does not match that experimentally determined previously[45].

Salt bridges increase kinetic stability of collagens

The anomalously high frequency of KGE and KGD triplets has previously been rationalized based on their ability to foster salt bridges, which presumably increases the thermodynamic stability of collagens[21]. However, it is well-established that triple-helical peptides containing an OGP triplet unfold at a temperature similar to or marginally higher than that containing either KGE or KGD[26]. We also observed this with triple-helical peptides composed of the GPKGDO, GPKGEO or GPOGPO motifs flanked by three GPO triplets (Table 1). In view of this, the KGE or KGD triplets do not offer any thermal stability advantage compared to an OGP triplet. Thus, the evolutionary implications of retaining KGE/KGD triplets at a frequency comparable to OGP is not explained thermodynamically.

Generally, salt bridges increase the thermodynamic stability of proteins[37,48,49]. However, those with limited[50,51] or even destabilizing effects[35,52] on stability also abound. The destabilization arises due to unfavorable desolvation energy of the interacting electrostatic charges competing against favorable Coulombic attraction. Given the evolutionary pressure to retain even destabilizing salt bridges, it has been postulated that they modulate folding kinetics rather than influence thermodynamic stability[48,53]. This has led to a general appreciation of the role of salt bridges in kinetics irrespective of the thermodynamic component. For example, a surface arginine-glutamate salt bridge in staphylococcal nuclease hinders denaturation by raising the kinetic barrier to unfolding by ~7 kcal/mol[54]. Importantly, a geometrically optimized surface salt bridge slows the unfolding rate of α-helical peptides while unfavorable geometry has the opposite effect[55,56]. Given that KGE and KGD form geometrically specific salt bridges but do not offer any thermodynamic advantage compared to the OGP triplet, we also suspected a kinetic role.

Due to their trimeric nature, the rate of collagen nucleation is concentration dependent while also limited by the slow rate of proline cis-trans isomerization ($k_{cis \rightarrow trans} \sim 0.1\ s^{-1}$)[57]. Consequently, deconvoluting the effect of salt bridges on refolding rate from the rate-limiting proline isomerization is challenging. Thus, we investigated the unfolding kinetics of collagen peptides containing KGE or KGD triplets via circular dichroism (CD), differential scanning calorimetry (DSC) and molecular dynamic (MD) simulations.

Salt bridge interactions include electrostatic and hydrogen-bonding components. At pH lower than the pKa of glutamate and aspartate sidechains, the carboxylates are protonated abrogating the electrostatic component. Thus, we monitored the isothermal unfolding of KGE, KGD and OGP peptides at acidic and neutral pH to understand how salt bridges contribute to kinetics (see methods). At pH 2.5, the KGE and KGD peptides unfold to ~80% of the initial value after incubation at 37 °C for 8 h, but no significant change in signal is observed at neutral pH (Fig. 5A). Importantly, the signal for the OGP peptide is largely independent of pH. Unfolding of both peptides at neutral pH is slow with less than 10% loss in signal over 8 h (Fig. 5B). However, at pH 2.5, both peptides unfold quickly with a half-life of 1.6 and 3.5 h (Table 1). This pH dependence of unfolding half-life suggests that the increased kinetic stability of KGE and KGD triple helices at neutral pH is primarily due to salt bridge interactions.

We further investigated the unfolding kinetics of native fibrillar collagens type II and V and network-forming collagen type IV. As shown in Fig. 5C, all three collagen types show pH-dependent unfolding kinetics. The unfolding traces could be fit to a biexponential kinetics with a faster and slower decay phase. The half-life of the faster unfolding phase is largely independent of pH as well as the type of collagen except type V where a two-fold difference is observed. In contrast, the half-life of the slower unfolding phase is pH-dependent

https://doi.org/10.1038/s41467-024-54046-y

Table 1 | Stability and kinetic parameters for the unfolding of model collagen peptides and native collagens

Technique	Parameter	Peptides[d]			Native collagens[e]		
		OGP[a]	KGD[b]	KGE[c]	Type II	Type IV	Type V
DSC	T_m (°C)	67.6 ± 2.5 (67.5 ± 2.4)	62.6 ± 1.9 (60.7 ± 2.1)	58.9 ± 2.1 (56.3 ± 2.0)	–	–	–
	E_a (kcal mol^{-1}) from individual fitting	59.7 ± 1.2 (59.5 ± 1.1)	70.8 ± 1.3 (68.1 ± 1.2)	73.3 ± 0.8 (70.6 ± 0.7)	–	–	–
	E_a (kcal mol^{-1}) from Arrhenius plot	58.8 ± 0.4 (59.1 ± 0.5)	70.4 ± 0.5 (68.9 ± 1.9)	71.7 ± 0.7 (71.8 ± 0.6)	–	–	–
	k (min^{-1}, 25 °C) from Arrhenius Plot	1.59e−06 (1.595e−06)	7.91e−07 (8.961e−07)	7.47e−07 (8.345e−07)	–	–	–
	Half-life (days, 25 °C)	302.9 (301.8)	608.8 (537.2)	644.2 (576.8)	–	–	–
CD	$k_{pH=2.5}$ (min^{-1}, 37 °C)	–	–	–	5.6e−04 ± 1.1e-4 (1.5e −02 ± 1.9e−03)	1.2e−03 ± 2.1e−4 (1.3e −02 ± 1.0e-03)	1.0e-03 ± 3.9e−4 (3.4e −02 ± 3.4e-03)
	$k_{pH=7.4}$ (min^{-1}, 37 °C)	–	–	–	2.9e−04 ± 5.5e−5 (1.3e −02 ± 2.4e-03)	2.9e-04 ± 4.4e-5 (1.3e −02 ± 1.9e-03)	2.8e−04 ± 5.2e-5 (1.4e −02 ± 2.4e−03)
	Half-life$_{pH=2.5}$ (hours, 37 °C)	–	–	–	20.8 (0.78)	9.5 (0.91)	11.4 (0.35)
	Half-life$_{pH=7.4}$ (hours, 37 °C)	–	–	–	39.3 (0.95)	39.5 (0.90)	41.9 (0.83)
MD	simulation time (μs)	199.0	–	210.4	–	–	–
	MFPT$_{off}$ (ms)	0.7 ± 0.1		50.0 ± 14.7 (20.2 ± 9.5)	–	–	–

[a]OGP = Ac-(GPO)$_3$GPOGPOG(GPO)$_3$-NH$_2$; [b]KGD = Ac-(GPO)$_3$GPKGDOG(GPO)$_3$-NH$_2$; [c]KGE = Ac-(GPO)$_3$GPKGEOG(GPO)$_3$-NH$_2$.
[d]neutral pH or in presence of 154 mM NaCl (parenthesis); MFPToff for KGD not reported as the implied timescale analysis did not converge.
[e]rate constants for faster unfolding phase are shown in parenthesis.

for all collagen subtypes. Collagen type IV and V unfold ~4-fold slower at pH 7.4 than at pH 2.5 while collagen type II unfolds two-fold slower, as judged from their half-lives.

We also investigated the unfolding of peptides KGD and KGE via DSC to determine the activation energy, which is the energy required to overcome the barrier separating the native and the unfolded states. We extracted the activation energy from a kinetic analysis of peptide unfolding endotherms obtained at different scan rates. The thermal-unfolding endotherms show a temperature shift of melting transition upon varying the scan rate (Supplementary Fig. 8). This suggests that a kinetic process controls unfolding. Kinetic control of collagen unfolding has also been observed by others[58]. In our case, fitting of the endotherms to a kinetic model provides an estimate of the unfolding activation energy (**methods**). This parameter has been previously correlated with the magnitude of the protein kinetic stability[59,60]. In comparison to the OGP peptide, the KGD and KGE peptides have a lower thermal stability but considerably higher activation energy (Table 1), both in the Arrhenius plot (Fig. 5D) and the endotherm fitting (Supplementary Fig. 8). These results indicate a clear difference in kinetic stability as inferred by the one order of magnitude change in the kinetic constant (k) and consequently the unfolding half-time of the peptides (Table 1).

In order to further confirm that the observed differences are indeed due to salt bridge interactions, we performed DSC measurements in the presence of physiological concentrations of NaCl (Supplementary Fig. 9). While the melting temperature (Tm) and activation energy are essentially unchanged for OGP in the presence of salt, a destabilizing effect is observed for both KGE and KGD (Supplementary Table 4). Based on these results, it would seem that the destabilization in the presence of salt would render the salt bridges ineffective as a folding code. However, this is not straightforward. These results, which are based on small triple-helical peptides containing just 2 salt bridges, cannot simply be extrapolated to natural collagens containing an average of 50 salt bridges due to several reasons discussed below.

While it is generally believed that high ionic strength screens electrostatic charges and should, theoretically, decrease protein stability, this is not always observed. The effect of ionic strength on protein stability and kinetics is strongly context-dependent. For example, the thermodynamic stability of small archaeal modifier protein 1 (SAMP1) from halophilic organism *Haloferax volcanii* increases with increase in ionic strength. Increased ionic strength also decreases the unfolding rate i.e. increases the kinetic stability of SAMP1[61]. Thus, in this case, high ionic strength appears to increase both thermodynamic and kinetic stability. In another report, while comparing the effect of salt on a cold-shock protein from mesophilic, thermophilic, and hyperthermophilic organisms, Dominy et al. found that the mesophilic protein is stabilized but those from a thermophilic and hyperthermophilic organism are destabilized in the presence of salt[62]. These examples underscore how physiological salt concentrations can have widely different, and sometimes opposing, effects depending on the context in which protein functions.

As noted previously, salt bridges in natural collagens predominantly occur in dense clusters we call salt bridge knots and many collagens contain several of these knots preferentially enriched around interruptions (Supplementary Table 3). While we currently do not understand the function of such highly localized clusters of salt bridges in collagens, a brief survey of salt bridge abundance and stability in thermophilic, hyperthermophilic and halophilic organisms offers a clue. Proteins from these organisms frequently contain complex salt bridges where one acidic or basic residue interacts with multiple oppositely charged partners. This cooperatively stabilizes proteins i.e. the total increase in stability is more than that obtained from a simple sum of individual salt bridges[32,33]. For example, the enzyme glutamate dehydrogenase from the hyperthermostable archaea *Pyrococcus furiosus* contains a network of complex salt

https://doi.org/10.1038/s41467-024-54046-y

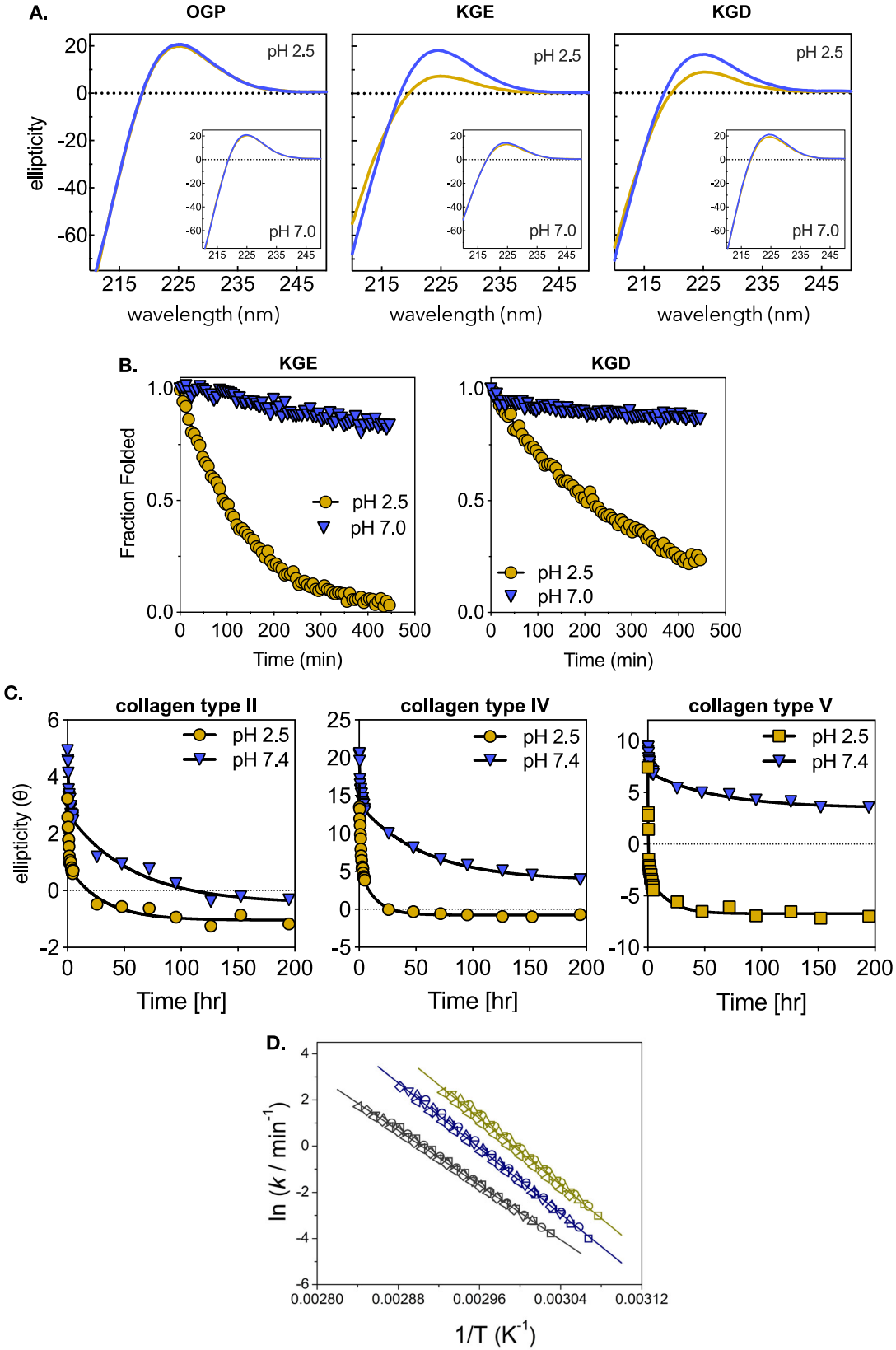

Fig. 5 | Salt bridges decrease the unfolding rate of collagens. A Decrease in the characteristic polyproline type II ellipticity signal of OGP, KGE, and KGD peptides after isothermal incubation at 37 °C. The initial CD spectra and after 8 h are shown in blue and yellow respectively. (B-C) Kinetics of unfolding of KGE and KGD peptides (**B**) and collagens type II, IV, and V at pH 2.5 (blue) and 7.0 (orange) (**C**). Residuals of the mono-exponential fit of collagen peptides and bi-exponential fit of native collagens are shown in Supplementary Fig. 7. **D** Kinetic stability comparison of OGP (gray), KGE (blue), and KGD (yellow) peptides analyzed by an Arrhenius plot derived from differential scanning calorimetry experiments performed at different scan rates. The line represents the best fit to the Arrhenius equation. Kinetic parameters are shown in Table 1. DSC endotherms and fitting to a two-state kinetic model are shown in Supplementary Fig. 8.

bridges, which is absent in the mesophilic homolog *Clostridium symbiosum*[63,64]. Similarly, halophilic proteins remain folded in intracellular and extracellular salt concentrations that approach 5 moles per liter. Bioinformatic analysis of halophilic proteins suggests an abundance of electrostatically charged amino acids capable of forming salt bridge networks[65,66]. It has also been argued that the increased abundance of salt bridges, especially on the solvent-exposed protein surface, contributes to their remarkable thermostability under extreme salt conditions[67,68]. We suspect that the high density of salt bridges in the surface-exposed salt bridge knots in collagens performs a function not very different from complex and networked salt bridges in extremophilic organisms. They likely compensate for the loss in stability due to the presence of physiological salt concentration.

It should also be noted that the effect of salt on native collagen stability and kinetics cannot be directly assessed due to their tendency to form large aggregates and ultimately hydrogels in presence of physiological salt concentrations. As a workaround, we are currently in the process of investigating this aspect using model triple-helical peptides containing salt bridge knots. We will report the outcome in due course. Nonetheless, we suspect that triple helices stabilized by such knots would be far less sensitive to ionic strength than the KGE and KGD peptides employed here.

The trend in kinetic stability of peptides as observed by CD and DSC experiments is further confirmed by molecular dynamics (MD) simulations. Starting from the previously published crystal structures of collagen peptides containing the KGD and KGE[26] or OGP[69] triplets, we conducted multiple all-atom simulations in parallel, each capable of sampling various conformational landscapes. Specifically, simulations were conducted for systems containing the triplets KGE, KGD, and OGP, alongside a simulation with the KGE triplet at a NaCl concentration of 150 mM (**methods**). In total, we accumulated over 750 μs of aggregated simulated time (Table 1). To facilitate unfolding processes within a reasonable timescale, all systems were maintained at 400 K. Each individual simulation ran for 100 ns, which prevented the observation of complete unfolding events but allowed us to sample partial fragments of the process. We therefore employed Markov state models (MSMs) to combine the information from the entire ensembles and analyze their collective behavior. MSMs have proven invaluable for characterizing kinetics, thermodynamics, and trajectories of processes like ligand-binding[70], folding[71], and protein-protein interactions[72] with atomic precision. We computed MSMs for all our systems, which allowed us to obtain approximate transition rates between the folded and unfolded states (methods and Supplementary Fig. 10). More specifically, in the context of MSMs, we define the mean first passage time (MFPT) as the average time required to transition from one state to another. We computed MFPTs for our simulations, thereby characterizing the average times required for the different systems to reach the unfolded state. The implied timescales for the OGP, KGE and KGE in presence of physiological salt concentrations converged but not KGD. Thus, KGD was excluded from further analysis. As indicated in Table 1, KGE triple-helical peptide unfolds approximately two orders of magnitude slower than those containing hydroxyproline or high salt concentration. To summarize, both theoretical and experimental investigations suggest that salt bridge interactions increase the kinetic stability of collagen triple-helical peptides more than those containing only proline and hydroxyproline. This provides a rationale for why Nature has retained the salt bridge forming triplets at anomalously high frequency in human collagens.

Mutations that disrupt salt bridges are associated with disease

Mutation of amino acids critical for protein folding and function causes disease. We investigated if mutations within or in the periphery of salt bridges in collagen are also associated with disease. It has previously been shown that a Gly → Ala mutation completely unfolds a region located 9 residues upstream towards the N-termini[73]. In a separate study, Gly → Ser mutation also induces unfolded monomer-like dynamics in the N-terminal residues[74]. Mutations also disrupt triple helical structure in the C-termini but to a more restricted extent than in the N-termini. For example, Gly → Ser mutation disrupts a salt bridge located two triplets towards the C-termini[75] but that located five triplets away shows a canonical triple helical conformation[74]. Thus, in effect, mutation of a glycine would minimally disrupt a salt bridge located three triplets towards the N-termini or two triplets in the opposite direction. We refer to this region over which a mutation can disrupt salt bridges and can potentially influence folding and stability as the salt bridge footprint (see methods). A salt bridge footprint is schematically shown in Supplementary Fig. 11A.

In order to understand if mutations within the salt bridge footprint are associated with disease, we searched the Clinvar[76], Leiden Open Variation Database (LOVD)[77] and Alport Syndrome Database[78] for pathogenic missense mutations due to single nucleotide change. We identified a total of 2294 mutations, of which 565 (~25%) occur within the salt bridge footprints (Supplementary Table 5). The salt bridge footprints in aligned sequences and the identified pathogenic missense mutations are shown in Supplementary Fig. 11B.

The most dramatic localization of mutations in the salt bridge footprint is observed in the α5 (38%) and α6 (52%) chains of collagen type IV and the α1 (45%) chain of collagen type VII. The association of salt bridge disruption to disease-causing mutations testifies to their biological function. It is anticipated that mutations would lower the kinetic stability of triple helices causing downstream effects such as change in fibrillogenesis, collagen turnover and transport and secretion to the extracellular matrix but the exact mechanism remains to be investigated.

Glycine mutations in the salt bridge footprint of collagen type I are associated with lethal phenotypes

Mutations in the triple-helical region of collagen type I cause osteogenesis imperfecta (OI), also known as brittle bone disease. Based on the mode of inheritance and clinical symptoms, Sillence et al. classified OI into four broad categories; OI1 presents moderate phenotype, OI2 is lethal causing stillbirth or perinatal death, OI3 causes most-severe but non-lethal phenotype while OI4 phenotype varies between 1 and 3[79]. A survey of the LOVD database suggests that 192 missense mutations present in the triple-helical domain of col1α1 and 164 in col1α2 cause lethal (OI2) or severe (OI3) phenotype (Supplementary Data 4). Of these, 54 mutations in the α1 chain (28%) and 40 mutations in α2 chain (24%) are present within the salt bridge footprint. Importantly, 55% (52 of 94) mutations in the salt bridge footprint of the α1 and α2 chains are lethal.

Previously, mutational hotspots in collagen type I that cause lethal OI have been mapped to two sequence stretches in the α1 chain (residue 869-1001 and 1088-1142) and 8 in α2 chain (residues 409-454, 541-592, 637-670, 712-727, 784-796, 847-901, 949-997 and 1027-1084)[80,81]. These lethal regions are schematically shown in Supplementary Fig. 12. It has previously been postulated that lethal regions in the collagen sequence correlate with the major ligand-binding regions of the triple-helical domain[82]. Presumably, mutations disrupt the interaction of collagen fibrils to other proteins, resulting in pathology. However, only 8 of the 54 lethal mutations in the α1 chain within the salt bridge footprint co-localize with the known lethal region. Our results suggest that in addition to protein-protein interactions, the molecular effects of mutations that potentially disrupt salt bridges could also be consequential in determining the severity of phenotype.

It should be noted that the mutations are not statistically more likely to appear in the footprint than elsewhere in the triple helices of native collagens. As shown in Supplementary Table 5, the salt bridge footprints account for ~43% of all triplets we analyzed. Yet, they contain only 33% of all pathogenic mutations. However, we would like to qualify this conclusion with a major caveat. A precondition for such a

statistical analysis is that the pathogenic or benign nature of missense mutation at every sequence position in all native collagens is known a priori. This is far from accomplished. We know the location of salt bridges and the footprint with precision but we don't know all mutations both within and outside the footprint that might contribute to disease. The current database of mutations is incomplete and pathogenic mutations at new residue positions are added as and when they are discovered based on clinical symptoms. Thus, we have only part of the information. For example, only one pathogenic missense mutation is known for collagen types XIII and XV, both of which are outside the footprint. In contrast, 1 and 3 mutations are known in collagens type XXIV and XXV, respectively, all of which are within the salt bridge footprint. Furthermore, there are 42 pathogenic mutations identified in the α1 chain of collagen type XI but it is an ABC heterotrimer composed of an α1, α2 and a3, which we cannot unambiguously align the chain. Thus, which proportion of the 42 mutations are within or outside the footprint is not clear to us. These numbers will change as and when more mutations are identified and so will the statistical analysis. Thus, in our view, the interpretation of the outcome of the statistical analyses at this stage is largely meaningless.

Despite the lack of statistical significance, the hypothesis that the disruption of the salt bridge and the resulting decrease in kinetic stability could explain the location-dependent phenotypic severity of natural collagens is intriguing. Remaining folded long enough is a precondition for the function of nearly all proteins. Thus, mutations that are detrimental to kinetic stability are likely to cause pathology. In our case, mutations that cause a loss in kinetic stability would induce local unfolding of the triple helix. While seemingly a minor event, this microunfolding is likely to cause a catastrophic cascade. It can render the triple helix more susceptible to cleavage by collagenases. It can also cause over-modification of the proline and lysine residues by prolyl hydroxylase and lysyl oxidase, respectively, considerably changing stability and downstream hierarchical organization into fibers, networks, or microfibrils as well altered recognition of other proteins. In light of this, we hypothesize that the phenotypic severity of a mutation could be determined by how strongly it alters the kinetic stability of the collagen triple helix. We are currently investigating this hypothesis using a combination of model collagen peptides and recombinantly expressed mutant collagen proteins.

Discussion

Collagen is an enigmatic protein from a folding perspective. Unlike globular and other coiled-coil proteins, the triple helix lacks a hydrophobic core and is primarily stabilized by interchain hydrogen bonds formed by the repetitive glycines. Amino acids in the Xaa and Yaa positions determine various aspects of folding and stability. Collagen α-chains are generally translated with a globular C-terminal pro-domain[83] that self-trimerizes and directs assembly of the correct triple-helical stoichiometry while also facilitating rate-limiting nucleation. Importantly, the pro-domains are cleaved before the nascently folded triple helices undergo further self-assembly into fibers, networks, and filaments. In a remarkable work, Leikina et al. have shown that collagen type I unfolds at body temperature[84]. The implication is that the nascent triple helices have unfavorable free energy and would start unfolding as soon as the pro-domains are cleaved. In order to explain how Nature circumvents this folding problem, it has been proposed that the ER-resident chaperone Hsp47 binds unique sites in triple helices. This interaction is believed to minimize local unfolding, which could be a precursor to global unfolding, and also offset the entropic cost for the stepwise propagation of a ~1000 amino acid long polypeptide[85,86]. However, this folding paradigm does not reconcile with the experimental observation that exogenously added Hsp47 does not significantly influence either the stability or the folding rate of collagen type I in vitro[87]. Several other biochemical observations are also not fully consistent with the Hsp47-chaperoned folding paradigm.

First, Hsp47 binding sites are predominantly located towards the N-terminal half of collagens type I, II, and III[12,20,88]. Given the C- to N-terminal propagation, the implication is that Hsp47 binds these collagens after half the triple-helix has already folded. Second, Hsp47 does not recognize collagens type XI, XIV, XXII, or does so only weakly[12]. This is also corroborated by our analysis of the distribution of potential Hsp47 recognition sites across the 28 human collagens (Supplementary Fig. 13A–C). We find that collagens type VIII, XIII, XIV, XV, XVIII, XIX, XXI, and XXV contain less than 3 Hsp47 binding sites even when low affinity sites are considered. Third, Hsp47 chaperoned folding runs counter to the general experimental observation that native collagens spontaneously refold in vitro in its absence. Fourth, the role of Hsp47 during late stages of collagen folding such as preventing premature lateral aggregation of folded triple-helices[89,90], transporting pro-domain cleaved triple-helices to the ER-Golgi boundary[91,92] and then sorting into large pro-collagen cargos[93] for timely secretion into the extracellular matrix is well-established. The experimental observations that Hsp47 ablation does not influence the secretion of collagen type VI microfibrils but only its lateral assembly supports this argument[12]. We also find that fibrillar collagens, which are composed of laterally assembled triple helices, contain the highest abundance of Hsp47 binding sites (Supplementary Fig. 13D). Finally, collagen-like protein domains of considerable length, and interruptions are also found in archaea, bacteria, eukarya (excluding human collagen orthologues) and viruses. To the best of our knowledge, there is currently no evidence for the presence of Hsp47-like proteins in organisms from these groups. Thus, how collagens and collagen-like proteins in these groups fold and then remain folded is an open question.

Salt bridges help globular proteins avoid unproductive folding pathways by stabilizing productive folding intermediates[94,95]. Here, we propose that the salt bridges also stabilize the propagating triple helix, in effect a folding intermediate, until it has reached a sufficient length to counter entropic disruption of further folding. This folding paradigm is a direct consequence of the geometric specificity of salt bridges and their ability to decrease the unfolding rate of triple helices. We find that the 28 human collagens contain 1553 lysine–aspartate/glutamate salt bridges with an average of 50 in each collagen subtype. We also find that the interaction between lysine and aspartate/glutamate residues is geometrically specific and that they increase the kinetic stability of model triple-helical peptides as well as native collagens. Of the 1553 salt bridges, 50% are localized in the C-terminal quarter of the triple-helical domains. This highly skewed concentration in the C-terminal quarter along with their ability to increase kinetic stability suggests that salt bridges act as electrostatic clamps to prevent local unfolding and, in the process, allow triple helices to reach a critical length from which the rest of the propagation can occur. Our observation that collagens with longer triple-helical domains or a greater number of structurally and energetically disruptive non-collagenous interruptions also have more salt bridges further corroborates this hypothesis. Additionally, the observation that interruptions are predominantly flanked by multiple closely spaced salt bridges, which we call salt bridge knots, adds further weight to it.

CD and DSC experiments suggest that the increased kinetic stability is a result of a raised kinetic barrier to unfolding. Importantly, this kinetic barrier is lower for the triple helices containing only proline and hydroxyproline than those containing salt bridges. These findings shed light on the counterintuitive observation that type I collagen is thermodynamically unstable, despite the abundance of OGP triplets. By raising the kinetic barrier, salt bridges delay collagen unfolding, allowing sufficient time for hierarchical self-organization into higher-order structures to occur. This organization can contribute to a packing force that would further prevent unfolding. This provides a rationale for why KGE and KGD triplets are anomalously frequent despite a thermal stability contribution comparable to OGP.

We also considered if the lysine-containing triplets could be anomalously frequent for reasons other than folding. Individual charged residues are pivotal to the assembly of collagen molecules into fibrils as well as recognition of other proteins. For example, oxidation of lysine to hydroxylysine is needed for inter-triple-helical cross-linking of fibers[96]. Although collagen sequences are very rich in lysines, only a small fraction of these are part of the KGE and KGD triplet. For instance, the α1 chain of type I collagen contains 37 lysines, of which only 7 are part of a KGE or KGD triplet. The remaining 24 are available for crosslinking of collagen fibrils. In light of this, we can speculate that the likelihood of lysines in KGE and KGD triplets to be involved in collagen crosslinking, and to be retained by Nature for this reason, is small. Similarly, collagen binding sites often contain lysine, aspartate, or glutamate residues (in particular in GxxGEx integrin recognition motifs). However, no binding motifs containing KGE/KGD triplets have been identified to our knowledge[97]. The prevalence of lysine and aspartate/glutamate close in sequence space as in KGE or KGD is what drives the formation of the many salt bridges in human collagens. Thus, we speculate that there is an evolutionary pressure to retain KGE or KGD motifs for reasons of folding, rather than for crosslinking or protein recognition sites.

We also considered if the OGP triplet, also found at anomalous frequency in human collagens, could also be part of the folding code. However, several arguments suggest otherwise. First, KGE and KGD triplets guide folding due to their ability to form interchain salt bridges. The OGP sidechains lack this capability. The proline and hydroxyproline residues pre-organize the main chain[98] and the latter also form water-mediated hydrogen-bonds[99] but they do not engage in specific interchain interactions. The salt bridge interaction is also constrained by both sequence and geometry. This specificity of interaction with respect to sequence and geometry is the underlying basis of the code that we refer to in the manuscript. Second, salt bridges slow the rate at which triple helices unfold. While OGP also slows the unfolding rate, the activation energy for the unfolding of OGP containing triple helices is significantly less than those containing salt bridges (Table 1). This is reflected in the 2-fold smaller half-life of OGP than KGE/KGD containing triple helices. We believe that a key reason for the lower kinetic stability imparted by OGP triplets is the lack of stabilizing interchain interactions that could hinder unfolding. Third, OGP triplets are not anomalously frequent in the region surrounding interruptions. In contrast, the increased kinetic stability and the specificity of interaction is the primary reasons why KGE and KGD-mediated salt bridges can also compensate for the deleterious effects of interruptions. This is also corroborated by the strong positive correlation observed between the number of interruptions and the number of salt bridges. Markedly, only a weak correlation exists between the interruptions and the number of OGP triplets. Finally, OGP triplets are binding sites for the Glycoprotein (GP) VI receptor[100], a key receptor for the activation and aggregation of platelets as well as the cell-lysis inhibitory receptor Leukocyte-associated immunoglobulin-like receptor-1[101]. As a result, it is difficult to appreciate if the observed prevalence and localization of OGP triplets in human collagen is due to their contributions to the folding/stability of the triple helix, to their biological role in triggering platelet activation, to the prevention of lysis when the cell is identified as self, or all of these functions. As a result, while OGP is crucial to the stability of the collagen triple helix, we cannot distinguish between its different roles and thus cannot draw conclusions based on their specific localization within the triple helix. This is contrary to KGE/KGD triplets that seldom appear in protein binding sites.

The salt bridge-assisted folding mechanism circumvents the need for a chaperone during the triple-helical propagation phase. This folding paradigm likely also explains how collagen-like proteins from archaea, bacteria, eukarya and viruses, with their stupendously long triple-helical domains and many interruptions, are able to successfully fold and remain folded. The wider implication is that evolution has converged on a similar mechanism to stabilize triple helices, making KGD the only triplet with anomalously high frequency across the three domains of life as well as viruses. Testifying to the genome-wide function of salt bridges, Mohs et al. have demonstrated that the collagen-like domain of bacterial protein *Streptococcus pyogenes* cell-surface protein Scl2 does not contain any hydroxyproline and that it is stabilized by salt bridges. This is deduced from a comparison of the melting temperature and molar residue ellipticity of recombinant Scl2 at pH 7 and pH 2.2[102]. Similarly, Ghosh et al. have shown that the protein EPclA from pathogenic enterohemorrhagic *Escherichia coli* strain O157:H7 is thermally more stable ($Tm = 42\,°C$) than the mammalian collagen, despite its much shorter sequence length. EPclA lacks hydroxylated prolines. Its remarkable stability is attributed to an abundance of proline in the Xaa position as well as to the roughly 22% electrostatically charged residues found within the sequences[18]. Our analysis revealed that the triplets containing threonine are also anomalously frequent in archaea, bacteria and viruses but the collagen-like domains in both Scl2 and EPclA do not contain an abundance of threonine-containing triplets. Thus, these collagen-like proteins appear to be stabilized by proline or electrostatic interactions or a combination of both.

The TGP triplet, anomalously frequent in viruses, has previously been incorporated in a collagen peptide. This peptide Ac-(Gly-Pro-Thr)$_{10}$-NH2 is a random coil in solution but glycosylation of all ten threonine residues with beta-D-galactose renders the peptide triple-helical[103]. This suggests that glycosylation of threonine is an alternative stabilizing mechanism to the hydroxylation of prolines. Organisms from all taxa possess the ability to glycosylate threonine. Whether this glycosylation occurs on the unfolded polypeptide or fully folded triple helices is currently unknown, to the best of our knowledge. Clearly, this will determine the degree to which TGP triplets contribute to folding and require further experimental investigation. We also noted that QGP triplet is anomalously frequent in archaea. A class of enzymes called Tissue Transglutaminase (tTG) has been shown to deamidate the glutamine in a QGP triplet in collagen type II to glutamic acid and facilitate inter-triple-helical crosslinking via imine linkage[104]. Homologous transglutaminases have also been identified in archaea[105]. Whether or not the transglutaminases in archaea post-translationally modify the glutamines in the QGP triplets remains to be investigated.

Our results concerning the role of salt bridges have additional important consequences for collagen folding. The three polypeptides of a collagen triple helices are staggered by one amino acid with respect to each other to optimize interchain hydrogen bonds and van der Waals packing. Given the enormous length of triple-helical domains and the diversity of interruptions, incorrect staggers of more than one amino acid are plausible. How collagens avoid such misfolded states is currently not understood. Our observation that incorrect stagger of α-chains reduces the number of salt bridges by an average of 50% across all collagen subtypes suggests a mechanism for how collagens might ensure correct stagger. Most revealingly, we find that mutations of glycine residues within or in the periphery of a salt bridge are associated with heritable diseases across several collagen types while such mutations in collagen type I frequently result in lethal phenotype. Previously, efforts to rationalize why mutation of some glycine residues in collagen type I causes a more severe phenotype than others using loss in thermal stability showed a poor correlation[106]. We can extrapolate from our results that the phenotypic severity is likely correlated to the loss in kinetic rather than thermodynamic stability. The structural consequence of mutation in a salt bridge footprint, how it affects folding kinetics, stability, fibrillogenesis, and ultimately transportation into the extracellular matrix, and why it overwhelmingly results in lethal phenotype is not clear to us. However, these questions gain significance in light of our observation that mutations in salt bridges generally result in lethal to severe

phenotypes, even when present outside the aforementioned lethal regions. Given the effect of salt bridges on collagen kinetic stability as demonstrated here, a systematic study to explore the correlation between loss in kinetic stability and the phenotypic severity due to a mutation is desirable in the future.

Methods

Identification of the collagen domains and interruptions

UniProtKB database was queried on 16.03.2023 for proteins containing the collagen or collagen-like sequences by applying taxonomic filters (archaea, bacteria, eukarya or virus) while filtering for sequences annotated by InterPro as containing collagen triple-helical repeats (ipr008160). The search terms used were "(taxonomy_id:2759) AND (xref:interpro-ipr008160)" for eukarya, "(taxonomy_id: 10239) AND (xref:interpro-ipr008160)" for viruses, "(taxonomy_id:2157) AND (xref:interpro-ipr008160)" for archaea, "(taxonomy_id:2) AND (xref:interpro-ipr008160)" for eubacteria. The 44 α chains of human collagen sequences were also obtained from UniProtKB by applying taxonomic filter 9606 and searching for gene names such as COL1A1, COL2A1, etc. A total of 59226 collagen sequences were downloaded from UniProtKB, which were split into smaller datasets based on the phyla they were recovered from. CD-HIT[107], a clustering algorithm to produce non-redundant dataset was used to cluster these collagen domains from each taxa. The sequences were clustered at 70% identity to remove most of the proteins with high sequence similarity and retain enough diversity. After the removal of duplicate sequences, collagen or collagen-like domains in the taxa-specific FASTA files were identified by searching for contiguous repeats of 6 or more Yaa-Gly-Xaa triplets. Given the 20 canonical amino acids and excluding Glycine at the Xaa and Yaa positions, a total of 361 residue pair combinations are theoretically possible in the Yaa-Gly-Xaa triplet. The number of each of the 361 possible Yaa-Gly-Xaa triplets in each collagenome was counted and their relative abundance was determined using Eq. 1 below,

$$rel.abundance = \frac{count\ of\ observed\ YGX\ triplets}{count\ of\ all\ observed\ triplets} \times 100 \quad (1)$$

The predicted distribution of the triplets was calculated based on the statistical model described previously[21]. We counted the observed instances of the different triplets and calculated Z-scores from the numerical difference in the observed and predicted count of triplets using the Z-score function in R package[108]. Motifs with Z-scores greater than 3 were considered outliers and thus anomalously frequent.

Interruptions in the human collagen domains and the collagen-like proteins in archaea, bacteria, eukarya, and viruses were identified essentially as described previously[24]. Briefly, interruptions were defined via their sequence length. For example, a deletion of either the Xaa or Yaa residue from a Yaa-Gly-Xaa triplet was denoted as G1G; a deletion of both Xaa and Yaa was denoted G0G; an insertion of n residues where $3 \le n \le 15$ was denoted GnG. Furthermore, motifs where glycine occupies the Xaa or Yaa position were identified as G2G interruptions. These motifs were extracted from the human and taxa-specific FASTA files containing the collagen and collagen-like domains, along with 7 amino acids on the N- and C-termini, to obtain the sequences of the interruptions together with the flanking P1-P4 triplets. Duplicate or redundant sequences were removed retaining only unique interruptions.

The amino acids on either side of the interruptions contain a total of four Yaa-Gly-Xaa triplets labeled here as P1 to P4 starting from the N-termini. In the case of G0G interruptions, manual inspection revealed false positives when allowing more than one glycine in the X or Y position of the P1 to P4 triplets. To circumvent this, we allowed only one glycine in the X or Y position which could be in the P1, P2, P3, or P4 position. Z-scores for the frequency of triplets surrounding interruptions were calculated as described in the case of the analysis of uninterrupted triple-helical domains. A list of taxa-specific interruptions is provided in the file Supplementary Data 2.

Determination of the chain alignment of native collagens

Despite stupendous advances in technology, obtaining the molecular structure of natural collagens has proven to be incredibly difficult due to their highly complex hierarchical architecture and the tendency to form insoluble aggregates upon extraction from native tissues. Thus, the alignment of collagens cannot be experimentally verified. Consequently, the current notion of the correct alignment in native collagens is a conjecture. However, in the case of some of the more widely studied collagens such as types I, II, III, IV VI, and IX, this conjecture is informed by sequence-based and experimental considerations. Below, we discuss these considerations for the aforementioned collagen subtypes. In the case of the remaining collagens, we relied on the optimization of the triple-helical domain as a criterion for obtaining the alignment. We also briefly discuss this below and present it in a tabular format in Supplementary Table 2.

Collagen type I. The alignment of collagen type I was previously deduced by investigating the interaction of an epitope within it with other proteins[45]. Briefly, the epitope could only recognize the proteins (von willebrand factor A3 domain and discoidin domain receptor 1 and 2) if presented in the correct alignment. The authors designed several alignments, of which, only one recognized all three proteins with the highest affinity. The authors presumed that this alignment is the correct one for collagen type I.

Collagens type II and III. Similar to collagen type I, the alignment of collagen type II and III was also deduced from a consideration of epitope alignment needed to recognize cell adhesion receptors α1β1 and α2β1 integrins[109] as well as discoidin domain receptors 1 and 2[110].

Collagen type IV. The alignment of collagen type IV [$(\alpha1)_2\alpha2$ heterotrimer] was deduced from the interaction of an epitope with α1β1 integrin[111]. This alignment was also deduced by Hohenester et al. via analysis of putative osteonectin binding sites in the $(\alpha1)_2\alpha2$ heterotrimer of collagen type IV[112]. This alignment was additionally confirmed by Parkin et al.[82], which was also the basis for aligning the $(\alpha5)_2\alpha6$ and α3α4α5 heterotrimers of collagen type IV.

Collagen type VI. The alignment of the α3α2α1 heterotrimer of collagen type VI was deduced by Ball et al. based on an epitope proposed to interact with integrin α2β1[113]. The alignment of the remaining α5α2α1 and α6α2α1 heterotrimers was based on optimization of the triple helical domain.

Collagen type IX. Here, we relied on the alignment proposed by Käpylä et al. based on the interaction of collagen type IX with the αI-domains of integrins α1β1, α2β1, α11β1, and α10β1[114]. Collagen type IX also contains four non-collagenous domains (NC1-NC4) interspersed between the collagen domains. Boudko et al. presented an alignment of these NC domains based on the localization of cysteine knots in NC1, NC2, and NC3 domain[115,116]. This further confirmed the alignment we used for our analysis.

Remaining collagen types. The alignment of all the remaining collagen subtypes was deduced by ensuring that the available Yaa-Gly-Xaa triplets on the three chains formed the longest triple-helical domain. It should be noted that all collagen types aligned in this fashion are homotrimeric. Thus, any interruptions present within the triple-helical domain are commensurate as defined previously by Bella et al.[24]. Consequently, the alignment of the chains as shown in Supplementary Fig. 1 does not contain any gaps, and the alignment before and after interruptions remains unchanged.

As noted before, the molecular structure of native collagens is difficult to obtain due to their highly complex and hierarchical nature and conformational flexibility. Consequently, these conjectures are perhaps the best we can do until further advances in experimental structural techniques reveal the molecular structure and the correct alignment of polypeptides in native collagens. It should also be noted that the alignments of collagen alpha-chains are unlike conventional multiple sequence alignments (MSAs). Here, they denote the alignment of three chains along the principal axis of the triple helix. Similarly, gaps in the sequence of triple-helical alignments, denoted by dash, also carry a different meaning. In a typical MSA, gaps account for insertions or deletions across related sequences. However, the dashes here account for incommensuration as defined previously by Bella et al.[24], and should not be construed as gaps in the traditional sense.

In some cases of heterotrimeric triple helices shown in Supplementary Fig. 1, all three chains may or may not carry interruptions. Moreover, the length of interruptions could also differ between the three chains. These cause changes in the offset before and after the interruption and are called incommensurate. In order to compensate for this incommensuration, gaps are introduced such that the three chains on either side are offset by one amino acid relative to each other. Structural consequences of incommensurate interruptions are currently not understood, as crystal structures of collagen triple helices containing such interruptions are not available. However, it has previously been proposed that such interruptions introduce a kink in the triple helix[117]. We speculate that the interruptions within the kink are likely to adopt semi-structured loop conformations, allowing the registration on either side to proceed in the canonical one-residue offset. However, this requires further experimental investigation.

Determination of the salt bridge footprint

As noted in the main text, mutation of glycine can potentially disrupt salt bridges located up to three triplets towards the N- and up to two triplets towards the C-termini. This region is termed the salt bridge footprint (Supplementary Fig. 11). Since mutation of lysine and aspartate/glutamate residues themselves can most obviously disrupt a salt bridge, these residues are also considered to be a part of the footprint. As noted previously, the i^{th} lysine of the leading and middle chains forms a salt bridge with the $i+2$ aspartate/glutamate of the middle and trailing chain, respectively. However, the i^{th} lysine of the trailing chains pairs with the $i+5$ aspartate/glutamate of the leading chain. As a consequence, the footprint for i to $i+5$ salt bridge is longer than the i to $i+2$.

Peptide synthesis

The Ac-$(GPO)_3$GPKGEO$(GPO)_3$-NH_2 (KGE), Ac-$(GPO)_3$GPKGDO$(GPO)_3$-NH_2 (KGD) and the Ac-$(GPO)_8$-NH_2 (OGP) peptides were synthesized on a TentaGel Rink Amide MBHA resin at a scale of 0.1 mmol using standard Fmoc-based solid-state peptide synthesis chemistry on a CEM microwave peptide synthesizer. The peptides were acetylated at the N-terminal and amidated at the C-terminal and cleaved from the resin using 95:2.5:2.5 volumetric mixture of trifluoroacetic acid (TFA), triisopropylsilane and miliQ-H_2O. Cleaved peptides were precipitated from TFA solution using dry-ice-cold diethyl ether, filtered under vacuum, re-dissolved in 95:5 water:acetonitrile (0.1% TFA), freeze-dried and stored at −20 °C. The cleaved peptides were purified using reverse-phase HPLC using a gradient of acetonitrile in water with 0.1% TFA. Multiple fraction corresponding to the main eluting peak were collected and analyzed by matrix-assisted laser desorption ionization mass spectroscopy. Fractions containing the desired peptide were pooled, flash-frozen in liquid nitrogen and lyophilized to obtain pure peptides used in all further experiments.

Determination of the unfolding rates via Circular Dichroism

The unfolding of peptides were determined to investigate the effect of electrostatic interactions on their kinetic stability. Peptides OGP, KGE, or KGD at pH 2.5 (aqueous HCl) or 7.0 (10 mM sodium phosphate buffer) at 0.4 mg/ml total peptide concentration were equilibrated in an Eppendorf Thermomixer at 37 °C and the characteristic CD maximum of polyproline type II helices at 225 nm monitored as a function of time until at least 80% of the initial signal has decayed. For measuring the spectra, the peptide solutions were transferred to a quartz cuvette (pathlength = 1 mm) pre-equilibrated at 37 °C via a Peltier temperature controller attached to a Jasco-815 CD spectropolarimeter and CD spectrographs recorded with a data pitch of 5 s and response time of 4 s at predetermined intervals. The average dead time between the transfer of the solutions to the cuvette and the recording of the spectra was 3 s. All spectra were recorded three times with three independent samples prepared from a common stock solution. The ellipticity of the peptide solutions was plotted as a function of wavelength without further data processing.

The unfolding rates were measured at 37 °C to match the experimental condition for monitoring the unfolding of native collagens described later. 0.4 mg/ml peptide solutions at pH 2.5 or 7.0 were incubated at 37 °C in an Eppendorf Thermomixer and the change in CD signal at 225 nm was monitored as a function of time. For each data point, the signal at 225 nm was averaged for 2 min. The samples were stored at 37 °C in an Eppendorf tube between measurements. Unfolding spectra were recorded thrice with samples independently prepared from a common stock solution. The raw data for samples were fitted to a one-phase decay model in GraphPad Prism according to Eq. 2 below and the data normalized using Eq. 3.

$$Y = (Y0 - plateau)e^{-kx} + plateau \tag{2}$$

$$normalized(Yi) = \frac{Yi - plateau}{Y0 - plateau} \tag{3}$$

In order to understand how electrostatic interactions influence the kinetic stability of native collagens, we monitored the unfolding of kinetics of native collagens type II (bovine tracheal cartilage; Sigma-Aldrich C1188), type IV (human placenta; Sigma-Aldrich C5533) and type V (human placenta; Sigma-Aldrich C3657) in aqueous buffer at pH 7.4 (200 mM sodium phosphate, pH 7.4 containing 0.5 M glycerol to prevent fibrillogenesis) and aqueous HCl at pH 2.5. 1 mg of each collagen was suspended in 1 ml of neutral or low pH buffer precooled in a ice bath and vortexed for 10 min. After orbital shaking overnight in a cold room, the suspensions were centrifuged at 2200 g for 10 min at 5 °C and the supernatants were used for further experiments. In a typical kinetic experiment, the native collagen samples stored at 5 °C were equilibrated to 37 °C for 10 min in an Eppendorf Thermomixer and then transferred to a quartz cuvette (pathlength = 1 mm) also maintained at 37 °C in the spectropolarimeter. The remaining experimental conditions for data acquisition were identical to those used for the collagen triple-helical peptides.

Determination of activation energy via DSC

Differential Scanning Calorimetry (DSC) measurements were collected in a VP-Capillary DSC (Malvern Panalytical). Samples were assayed at 0.4 mg/mL in buffer 10 mM sodium phosphate pH 7.0, and varying the scan rate from 0.5 °C/min to 3 °C/min. For experiments to determine the effect of salt on the kinetic stability, thermal unfolding was measured in buffer 10 mM sodium phosphate pH 7.0 and 154 mM NaCl. All scans were collected after exhaustive dialysis and buffer degassing. In all cases, proper instrument equilibration was reached by running at

least 2 buffer-buffer scans before sample-buffer experiments. The last buffer-buffer scan was then used to subtract the signal from each peptide-buffer scan in order to perform all thermodynamic analysis.

Calorimetric transitions were adequately described by the two-state kinetics model (N → F) where N is the native peptide and F is the final state[118]. The kinetic conversion from N to F is described by a first-order rate constant (k) changing with temperature according to the Arrhenius equation (Eq. 4):

$$k = exp\left[-\frac{E_{act}}{R}\left(\frac{1}{T}-\frac{1}{T^*}\right)\right] \quad (4)$$

where T^* is the temperature at which the kinetic constant $k = 1\,min^{-1}$ and E_{act} is the activation energy between the native and the transition states that describe the unfolding process[60]. Here, E_{act} was used to compare the kinetic stability among peptides. Then, the apparent heat capacity which describes the endotherm is given by (Eq. 5):

$$C_P^{APP} = \frac{\Delta HE_{act}}{RT_m^2} exp(x) exp[-exp(x)]; x = \frac{E_{act}}{RT_m^2}(T - T_m) \quad (5)$$

where T is the temperature and ΔH is the unfolding enthalpy. The E_{act} was also obtained from the slope of Arrhenius plots, i.e. ln k vs. 1/T as described before[119]. The data and associated fits are presented in Supplementary Data 5.

Determination of unfolding times via MD simulations

The simulation systems were prepared taking the X-ray structures as protein templates (PDB code 3T4F and 3U29 for KGE and KGD[26], respectively). The molecules were modeled to 15 residues, with four residues at the N-terminus before the triplet and 8 after (Supplementary Fig. 10). Systems were solvated and ions were added to neutralization, except for the KGE system which was modeled at high concentration (150 mM NaCl). The systems were produced in all cases using the software HTMD[120] and had ca. 34,000 atoms, from which 33,500 corresponded to water molecules. All systems were minimized, equilibrated, and run using ACEMD[121] and amberff14SB[122] as forcefield at 400 K. with the TIP3P water model at 400 K for one replica during 2 ns. For the MD production, we ran an intelligent adaptive sampling scheme that performs the simulations in successive epochs by analyzing them with Markov state models (MSMs), starting with 10 generators for the first epoch[123]. The simulations ran with a multi-state integrator using a time-step of 4 fs in an NVT ensemble using the Langevin thermostat. The simulations ran using the Particle Mesh Ewald Method[124] with a cutoff of 9 Å for the van der Waals interactions and real-space electrostatic interactions. The metric used during the adaptive runs for the MSM analysis was the protein alpha carbon atoms. The analysis was performed with the software HTMD. The adaptive scheme ran in all cases until sufficient sampling was detected. In particular, 29%, 26%, and 19% of the trajectories had visited the unfolded state obtained in the MSMs analyses for KGE, GPO, and KGE-ions, respectively. A summary table is included in the supporting information detailing the aforementioned aspects of MD simulations (Supplementary Table 6).

Markov state modeling analysis

Markov state modeling proceeds from the discretization of the conformational space and the description of the dynamics of the system of study as a sequence of transitions between the discretized clusters. A properly discretized MSM shows converging timescales with high probability of transition among kinetically similar states, and a lower probability between kinetically separated states. From this model, the pathways and kinetic rates between distinct conformations may be derived[3]. Here, we ran 2104, 1486, 1990, and 2014 simulations of 100 ns each for the KGD, KGE, GPO, and KGE + ions systems, respectively. The trajectories were projected using Euclidean distances among all alpha carbon atoms as a metric. We projected the multidimensional data onto its slow-order parameters using TICA (time-lagged independent component analysis)[125]. We projected the data into the minimum number of TICA dimensions that provided a converged implied timescale, in particular, we used 2, 1, 1, and 3 TICA dimensions, respectively. 1000 clusters were computed using the mini-batch k-means algorithm[126] in all cases. The clusters were lumped together into 2 macrostates by the PCCA algorithm[127], using a lag time of 40 ns in all cases. First-order kinetics are derived from mean-first passage times (MFPT)[128]. These analyses were performed with HTMD[120]. We estimated errors for the unfolding times using a bootstrapping technique. For this, we performed 10 independent runs in which 20% of the trajectories were randomly eliminated and a new MSM was built after re-clustering.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets, including taxa-specific list of collagen-like domains, list of interruptions, data associated with DSC and CD experiments and statistical analysis are provided as supplementary data with this paper. The dataset associated with molecular dynamic simulations including the trajectories is available upon request from the corresponding author. Source data are provided as a Source Data file with this paper.

References

1. Ricard-Blum, S. The Collagen family. *Cold Spring Harb. Perspect. Biol.* **3**, 1–19 (2011).
2. Bella, J., Eaton, M., Brodsky, B. & Berman, H. M. Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science.* **266**, 75–81 (1994).
3. Park, S., Klein, T. E. & Pande, V. S. Folding and misfolding of the collagen triple helix: Markov analysis of molecular dynamics simulations. *Biophys. J.* **93**, 4108–4115 (2007).
4. Beck, K. et al. Destabilization of osteogenesis imperfecta collagen-like model peptides correlates with the identity of the residue replacing glycine. *Proc. Natl Acad. Sci. Usa.* **97**, 4273–4278 (2000).
5. Myllyharju, J. & Kivirikko, K. I. Collagens and collagen-related diseases. *Ann. Med.* **33**, 7–21 (2001).
6. Engel, J. & Prockop, D. J. The zipper-like folding of collagen triple helices and the effects of mutations that disrupt the zipper. *Annu. Rev. Biophys. Biophys. Chem.* **20**, 137–152 (1991).
7. Sun, X. et al. A natural interruption displays higher global stability and local conformational flexibility than a similar gly mutation sequence in collagen mimic peptides. *Biochemistry* **54**, 6106–6113 (2015).
8. Kurkinen, M., Taylor, A., Garrels, J. I. & Hogan, B. L. M. Cell surface-associated proteins which bind native type IV collagen or gelatin. *J. Biol. Chem.* **259**, 5915–5922 (1984).
9. Yagi-Utsumi, M. et al. NMR and mutational identification of the collagen-binding site of the Chaperone Hsp47. *PLoS One* **7**, 5–10 (2012).
10. Widmera, C. et al. Molecular basis for the action of the collagen-specific chaperone Hsp47/SERPINH1 and its structure-specific client recognition. *Proc. Natl Acad. Sci.* **109**, 13243–13247 (2012).
11. Koide, T. et al. Specific recognition of the collagen triple helix by chaperone HSP47: II. The HSP47-binding structural motif in collagens and related proteins. *J. Biol. Chem.* **281**, 11177–11185 (2006).
12. Köhler, A. et al. New specific HSP47 functions in collagen subfamily chaperoning. *FASEB J.* **34**, 12040–12052 (2020).
13. 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).

14. Linden, T. A. & King, N. Widespread distribution of collagens and collagen-associated domains in eukaryotes. *bioRxiv* 2021.10.08.463732 (2021).
15. Qiu, Y., Zhai, C., Chen, L., Liu, X. & Yeo, J. Current insights on the diverse structures and functions in bacterial collagen-like proteins. *ACS Biomater. Sci. Eng.* **9**, 3778–3795 (2023).
16. 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).
17. Zairi, M., Stiege, A. C., Nhiri, N., Jacquet, E. & Tavares, P. The collagen-like protein gp12 is a temperature-dependent reversible binder of SPP1 viral capsids. *J. Biol. Chem.* **289**, 27169–27181 (2014).
18. Ghosh, N. et al. Collagen-like proteins in pathogenic E. coli strains. *PLoS One* **7**, (2012).
19. Xu, Y., Keene, D. R., Bujnicki, J. M., Höök, M. & Lukomski, S. Streptococcal Scl1 and Scl2 proteins form collagen-like triple helices. *J. Biol. Chem.* **277**, 27312–27318 (2002).
20. Cai, H. et al. Identification of HSP47 binding site on native collagen and its implications for the development of HSP47 inhibitors. *Biomolecules* **11**, 983 (2021).
21. Stability, T., Persikov, A. V., Ramshaw, J. A. M. M., Kirkpatrick, A. & Brodsky, B. Electrostatic interactions involving lysine make major contributions to collagen triple-helix stability. *Biochemistry* **44**, 1414–1422 (2005).
22. Bella, J., Liu, J., Kramer, R., Brodsky, B. & Berman, H. M. Conformational effects of Gly-X-Gly interruptions in the collagen triple helix. *J. Mol. Biol.* **362**, 298–311 (2006).
23. Mohs, A., Popiel, M., Li, Y., Baum, J. & Brodsky, B. Conformational features of a natural break in the type IV collagen Gly-X-Y repeat. *J. Biol. Chem.* **281**, 17197–17202 (2006).
24. Bella, J. A first census of collagen interruptions: Collagen's own stutters and stammers. *J. Struct. Biol.* **186**, 438–450 (2014).
25. Persikov, A. V., Ramshaw, J. A. M., Kirkpatrick, A. & Brodsky, B. Amino acid propensities for the collagen triple-helix. *Biochemistry* **39**, 14960–14967 (2000).
26. 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).
27. Daubenspeck, J. M. et al. Novel oligosaccharide side chains of the collagen-like region of BclA, the major glycoprotein of the Bacillus anthracis exosporium. *J. Biol. Chem.* **279**, 30945–30953 (2004).
28. Mann, K. et al. Glycosylated threonine but not 4-hydroxyproline dominates the triple helix stabilizing positions in the sequence of a hydrothermal vent worm cuticle collagen. *J. Mol. Biol.* **261**, 255–266 (1996).
29. Xu, T., Zhou, C. Z., Xiao, J. & Liu, J. Unique conformation in a natural interruption sequence of Type XIX collagen revealed by its high-resolution crystal structure. *Biochemistry* **57**, 1087–1095 (2018).
30. Hwang, E. S. & Brodsky, B. Folding delay and structural perturbations caused by type IV collagen natural interruptions and nearby Gly missense mutations. *J. Biol. Chem.* **287**, 4368–4375 (2012).
31. DiChiara, A. S. et al. A cysteine-based molecular code informs collagen C-propeptide assembly. *Nat. Commun.* **9**, (2018).
32. Musafia, B., Buchner, V. & Arad, D. Complex salt bridges in proteins: Statistical analysis of structure and function. *J. Mol. Biol.* **254**, 761–770 (1995).
33. Olson, C. A., Spek, E. J., Shi, Z., Vologodskii, A. & Kallenbach, N. R. Cooperative helix stabilization by complex Arg-Glu salt bridges. *Proteins* **44**, 123–132 (2001).
34. Gvritishvili, A. G., Gribenko, A. V. & Makhatadze, G. I. Cooperativity of complex salt bridges. *Protein Sci.* **17**, 1285–1290 (2008).
35. Phelan, P. et al. Salt bridges destabilize a leucine zipper designed for maximized ion pairing between helices. *Biochemistry* **41**, 2998–3008 (2002).
36. Hendsch, Z. S. & Tidor, B. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.* **3**, 211–226 (1994).
37. Kumar, S. & Nussinov, R. Salt bridge stability in monomeric proteins. *J. Mol. Biol.* **293**, 1241–1255 (1999).
38. Donald, J. E., Kulp, D. W. & DeGrado, W. F. Salt bridges: Geometrically specific, designable interactions. *Proteins Struct. Funct. Bioinforma.* **79**, 898–915 (2011).
39. Pal, R. et al. Syn vs anti carboxylic acids in hybrid peptides: experimental and theoretical charge density and chemical bonding analysis. *J. Phys. Chem. A* **122**, 3665–3679 (2018).
40. 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).
41. Gauba, V. & Hartgerink, J. D. Self-assembled heterotrimeric collagen triple helices directed through electrostatic interactions. *J. Am. Chem. Soc.* **129**, 2683–2690 (2007).
42. Gurry, T., Nerenberg, P. S. & Stultz, C. M. The contribution of interchain salt bridges to triple-helical stability in collagen. *Biophys. J.* **98**, 2634–2643 (2010).
43. Kramer, R. Z. et al. Staggered molecular packing in crystals of a collagen-like peptide with a single charged pair. *J. Mol. Biol.* **301**, 1191–1205 (2000).
44. Zheng, H. et al. How electrostatic networks modulate specificity and stability of collagen. *Proc. Natl Acad. Sci.* **115**, 6207–6212 (2018).
45. Jalan, A. A. et al. Chain alignment of collagen I deciphered using computationally designed heterotrimers. *Nat. Chem. Biol.* **16**, 423–429 (2020).
46. Fallas, J. A., Gauba, V. & Hartgerink, J. D. Solution structure of an ABC collagen heterotrimer reveals a single-register helix stabilized by electrostatic interactions. *J. Biol. Chem.* **284**, 26851–26859 (2009).
47. Obarska-Kosinska, A., Rennekamp, B., Ünal, A. & Gräter, F. ColBuilder: A server to build collagen fibril models. *Biophys. J.* **120**, 3544–3549 (2021).
48. Kursula, I., Partanen, S., Lambeir, A. M. & Wierenga, R. K. The importance of the conserved Arg191-Asp227 salt bridge of triosephosphate isomerase for folding, stability, and catalysis. *FEBS Lett.* **518**, 39–42 (2002).
49. Anderson, D. E., Becktel, W. J. & Dahlquist, F. W. pH-induced denaturation of proteins: a single salt bridge contributes 3–5 kcal/mol to the free energy of folding of T4 Lysozyme. *Biochemistry* **29**, 2403–2408 (1990).
50. Sali, D., Bycroft, M. & Fersht, A. R. Surface electrostatic interactions contribute little to stability of barnase. *J. Mol. Biol.* **220**, 779–788 (1991).
51. Hong, Z., Ahmed, Z. & Asher, S. A. Circular dichroism and ultraviolet resonance raman indicate little Arg-Glu side chain α-helix peptide stabilization. *J. Phys. Chem. B* **115**, 4234–4243 (2011).
52. Carey, D. W., Schildbach, J. F. & Sauer, R. T. Are buried salt bridges important for protein stability and conformational specificity? *Nat. Struct. Biol.* **2**, 122–128 (1995).
53. Andersson, H. S. et al. The α-defensin salt-bridge induces backbone stability to facilitate folding and confer proteolytic resistance. *Amino Acids* **43**, 1471–1483 (2012).
54. Gruia, A. D., Fischer, S. & Smith, J. C. Molecular dynamics simulation reveals a surface salt bridge forming a kinetic trap in unfolding of truncated Staphylococcal nuclease. *Proteins Struct. Funct. Genet.* **50**, 507–515 (2003).
55. Meuzelaar, H. et al. Solvent-exposed salt bridges influence the kinetics of α-helix folding and unfolding. *J. Phys. Chem. Lett.* **5**, 900–904 (2014).
56. Meuzelaar, H., Vreede, J. & Woutersen, S. Influence of Glu/Arg, Asp/Arg, and Glu/Lys salt bridges on α-helical stability and folding kinetics. *Biophys. J.* **110**, 2328–2341 (2016).

57. Buevich, A. V., Dai, Q. H., Liu, X., Brodsky, B. & Baum, J. Site-specific NMR monitoring of cis-trans isomerization in the folding of the proline-rich collagen triple helix. *Biochemistry* **39**, 4299–4308 (2000).
58. Mizuno, K. et al. Kinetic hysteresis in collagen folding. *Biophys. J.* **98**, 3004–3014 (2010).
59. Sanchez-Ruiz, J. M. Protein kinetic stability. *Biophys. Chem.* **148**, 1–15 (2010).
60. Romero-Romero, S., Costas, M., Rodríguez-Romero, A. & Alejandro Fernández-Velasco, D. Reversibility and two state behaviour in the thermal unfolding of oligomeric TIM barrel proteins. *Phys. Chem. Chem. Phys.* **17**, 20699–20714 (2015).
61. Mizukami, T., Bedford, J. T., Liao, S. H., Greene, L. H. & Roder, H. Effects of ionic strength on the folding and stability of SAMP1, a ubiquitin-like halophilic protein. *Biophys. J.* **121**, 552–564 (2022).
62. Dominy, B. N., Perl, D., Schmid, F. X. & Brooks, C. L. The effects of ionic strength on protein stability: The cold shock protein family. *J. Mol. Biol.* **319**, 541–554 (2002).
63. Yip, K. S. P. et al. The structure of Pyrococcus furiosus glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure* **3**, 1147–1158 (1995).
64. Rice, D. W. et al. Insights into the molecular basis of thermal stability from the structure determination of Pyrococcus furiosus glutamate dehydrogenase. *FEMS Microbiol. Rev.* **18**, 105–117 (1996).
65. Lanyi, J. K. Salt dependent properties of proteins from extremely halophilic bacteria. *Bacteriol. Rev.* **38**, 272–290 (1974).
66. Reistad, R. On the composition and nature of the bulk protein of extremely halophilic bacteria. *Arch. Microbiol.* **71**, 353–360 (1970).
67. Dym, O., Mevarech, M. & Sussman, J. L. Structural features that stabilize halophilic malate dehydrogenase from an archaebacterium. *Science.* **267**, 1344–1346 (1995).
68. Nayek, A. et al. Salt-bridge energetics in halophilic proteins. *PLoS One* **9**, 1–11 (2014).
69. Okuyama, K., Miyama, K., Mizuno, K. & Bächinger, H. P. Crystal structure of (Gly-Pro-Hyp)9: Implications for the collagen molecular model. *Biopolymers* **97**, 607–616 (2012).
70. Kröger, P., Shanmugaratnam, S., Ferruz, N., Schweimer, K. & Höcker, B. A comprehensive binding study illustrates ligand recognition in the periplasmic binding protein PotF. *Structure* **29**, 433–443.e4 (2021).
71. Chodera, J. D., Swope, W. C., Pitera, J. W. & Dill, K. A. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.* **5**, 1214–1226 (2006).
72. Plattner, N., Doerr, S., De Fabritiis, G. & Noé, F. Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **9**, 1005–1011 (2017).
73. Xiao, J., Cheng, H., Silva, T., Baum, J. & Brodsky, B. Osteogenesis imperfecta missense mutations in collagen: Structural consequences of a glycine to alanine replacement at a highly charged site. *Biochemistry* **50**, 10771–10780 (2011).
74. Liu, X., Kim, S., Dai, Q. H., Brodsky, B. & Baum, J. Nuclear magnetic resonance shows asymmetric loss of triple helix in peptides modeling a collagen mutation in brittle bone disease. *Biochemistry* **37**, 15528–15533 (1998).
75. Xu, K., Nowak, I., Kirchner, M. & Xu, Y. Recombinant collagen studies link the severe conformational changes induced by osteogenesis imperfecta mutations to the disruption of a set of interchain salt bridges. *J. Biol. Chem.* **283**, 34337–34344 (2008).
76. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
77. Fokkema, I. F. A. C. et al. LOVD v.2.0: The next generation in gene variant databases. *Hum. Mutat.* **32**, 557–563 (2011).
78. Crockett, D. K. et al. The Alport syndrome COL4A5 variant database. *Hum. Mutat.* **31**, E1652–E1657 (2010).
79. Sillence, D. O., Senn, A. & Danks, D. M. Genetic heterogeneity in osteogenesis imperfecta. *J. Med. Genet.* **16**, 101–116 (1979).
80. Sałacińska, K. et al. Novel Mutations Within Collagen Alpha1(I) and Alpha2(I) Ligand-Binding Sites, Broadening the Spectrum of Osteogenesis Imperfecta – Current Insights Into Collagen Type I Lethal Regions. *Front. Genet.* **12**, 692978 (2021).
81. Marini, J. C. et al. Consortium for osteogenesis imperfecta mutations in the helical domain of type I collagen: Regions rich in lethal mutations align with collagen binding sites for integrins and proteoglycans. *Hum. Mutat.* **28**, 209–221 (2007).
82. Parkin, J. Des et al. Mapping structural landmarks, ligand binding sites, and missense mutations to the collagen IV heterotrimers predicts major functional domains, novel interactions, and variation in phenotypes in inherited diseases affecting basement membranes. *Hum. Mutat.* **32**, 127–143 (2011).
83. Bellamy, G. & Bornstein, P. Evidence for procollagen, a biosynthetic precursors of collagen. *Proc. Natl Acad. Sci. USA.* **68**, 1138–1142 (1971).
84. Leikina, E., Mertts, M. V., Kuznetsova, N. & Leikin, S. Type I collagen is thermally unstable at body temperature. *Proc. Natl Acad. Sci.* **99**, 1314–1318 (2002).
85. Makareeva, E. & Leikin, S. Procollagen triple helix assembly: An unconventional chaperone-assisted polding paradigm. *PLoS One* **2**, e1029 (2007).
86. Koide, T. et al. Specific recognition of the collagen triple helix by chaperone HSP47: Minimal structural requirement and spatial molecular orientation. *J. Biol. Chem.* **281**, 3432–3438 (2006).
87. Macdonald, J. R. & Bächinger, H. P. HSP47 binds cooperatively to triple Helical Type I collagen but has little effect on the thermal stability or rate of refolding. *J. Biol. Chem.* **276**, 25399–25403 (2001).
88. Abraham, E. T. et al. Collagen's primary structure determines collagen:HSP47 complex stoichiometry. *J. Biol. Chem.* **297**, 101169 (2021).
89. Ishida, Y. et al. Type I collagen in Hsp47-null cells is aggregated in endoplasmic reticulum and deficient in N-Propeptide processing and Fibrillogenesis. *Mol. Biol. Cell* **16**, 5356–5372 (2005).
90. Matsuoka, Y. et al. Insufficient folding of type IV collagen and formation of abnormal basement membrane-like structure in embryoid bodies derived from Hsp47-null embryonic stem cells. *Mol. Biol. Cell* **15**, 4467–4475 (2004).
91. Nakai, A., Satoh, M., Hirayoshi, K. & Nagata, K. Involvement of the stress protein HSP47 in procollagen processing in the endoplasmic reticulum. *J. Cell Biol.* **117**, 903–914 (1992).
92. Satoh, M., Hirayoshi, K., Yokota, S. I., Hosokawa, N. & Nagata, K. Intracellular interaction of collagen-specific stress protein HSP47 with newly synthesized procollagen. *J. Cell Biol.* **133**, 469–483 (1996).
93. Ishikawa, Y., Ito, S., Nagata, K., Sakai, L. Y. & Bächinger, H. P. Intracellular mechanisms of molecular recognition and sorting for transport of large extracellular matrix molecules. *Proc. Natl Acad. Sci. USA.* **113**, E6036–E6044 (2016).
94. Ibarra-Molero, B., Zitzewitz, J. A. & Matthews, C. R. Salt-bridges can stabilize but do not accelerate the folding of the Homodimeric coiled-coil peptide GCN4-p1. *J. Mol. Biol.* **336**, 989–996 (2004).
95. Stoycheva, A. D., Onuchic, J. N. & Brooks, C. L. Effect of gatekeepers on the early folding kinetics of a model β-barrel protein. *J. Chem. Phys.* **119**, 5722–5729 (2003).
96. Eyre, D. R., Paz, M. A. & Gallop, P. M. Cross-linking in collagen and elastin. *Annu. Rev. Biochem.* **53**, 717–748 (1983).

97. Farndale, R. W. Collagen-binding proteins: Insights from the Collagen toolkits. *Essays Biochem.* **63**, 337–348 (2019).
98. Shoulders, M. D., Satyshur, K. A., Forest, K. T., Raines, R. T. & Raines, P. R. Stereoelectronic and steric effects in side chains preorganize a protein main chain. *Proc. Natl Acad. Sci. Usa.* **107**, 559–564 (2010).
99. Vitagliano, L., Berisio, R., Mazzarella, L. & Zagari, A. Structural bases of collagen stabilization induced by proline hydroxylation. *Biopolymers* **58**, 459–464 (2001).
100. Miura, Y., Takahashi, T., Jung, S. M. & Moroi, M. Analysis of the interaction of platelet collagen receptor glycoprotein VI (GPVI) with collagen: A dimeric form of GPVI, but not the monomeric form, shows affinity to fibrous collagen. *J. Biol. Chem.* **277**, 46197–46204 (2002).
101. Brondijk, T. H. C. et al. Crystal structure and collagen-binding site of immune inhibitory receptor LAIR-1: Unexpected implications for collagen binding by platelet receptor GPVI. *Blood* **115**, 1364–1373 (2010).
102. Mohs, A. et al. Mechanism of stabilization of a bacterial collagen triple helix in the absence of hydroxyproline. *J. Biol. Chem.* **282**, 29757–29765 (2007).
103. Bann, J. G., Peyton, D. H. & Bächinger, H. P. Sweet is stable: Glycosylation stabilizes collagen. *FEBS Lett.* **473**, 237–240 (2000).
104. Dzhambazov, B., Lindh, I., Engström, Å. & Holmdahl, R. Tissue transglutaminase enhances collagen type II-induced arthritis and modifies the immunodominant T-cell epitope CII260-270. *Eur. J. Immunol.* **39**, 2412–2423 (2009).
105. Makarova, K. S., Aravind, L. & Koonin, E. V. A superfamily of archaeal, bacterial, and eukaryotic proteins homologous to animal transglutaminases. *Protein Sci.* **8**, 1714–1719 (1999).
106. Makareeva, E. et al. Structural heterogeneity of type I collagen triple helix and its role in osteogenesis imperfecta. *J. Biol. Chem.* **283**, 4787–4798 (2008).
107. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics* **17**, 282–283 (2011).
108. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, https://www.R-project.org/ (Vienna Austria, 2023).
109. Xu, Y. et al. Multiple binding sites in collagen type I for the integrins α1β1 and α2β1. *J. Biol. Chem.* **275**, 38981–38989 (2000).
110. Xu, H. et al. Collagen binding specificity of the discoidin domain receptors: Binding sites on collagens II and III and molecular determinants for collagen IV recognition by DDR1. *Matrix Biol.* **30**, 16–26 (2011).
111. Golbik, R., Eble, J. A., Ries, A. & Kühn, K. The spatial orientation of the essential amino acid residues arginine and aspartate within the α1β1 integrin recognition site of collagen IV has been resolved using fluorescence resonance energy transfer. *J. Mol. Biol.* **297**, 501–509 (2000).
112. Hohenester, E., Sasaki, T., Giudici, C., Farndale, R. W. & Bächinger, H. P. Structural basis of sequence-specific collagen recognition by SPARC. *Proc. Natl Acad. Sci. USA.* **105**, 18273–18277 (2008).
113. Ball, S., Bella, J., Kielty, C. & Shuttleworth, A. Structural basis of type VI collagen dimer formation. *J. Biol. Chem.* **278**, 15326–15332 (2003).
114. Käpylä, J. et al. The fibril-associated collagen IX provides a novel mechanism for cell adhesion to cartilaginous matrix. *J. Biol. Chem.* **279**, 51677–51687 (2004).
115. Boudko, S. P. et al. The NC2 domain of collagen IX provides chain selection and heterotrimerization. *J. Biol. Chem.* **285**, 23721–23731 (2010).
116. Boudko, S. P. & Bächinger, H. P. Structural insight for chain selection and stagger control in collagen. *Sci. Rep.* **6**, 1–8 (2016).
117. Kilchherr, E., Hofmann, H., Steigemann, W. & Engel, J. Structural model of the collagen-like region of C1q comprising the kink region and the fibre-like packing of the six triple helices. *J. Mol. Biol.* **186**, 403–415 (1985).
118. Sanchez-Ruiz, J. M. Theoretical analysis of Lumry-Eyring models in differential scanning calorimetry. *Biophys. J.* **61**, 921–935 (1992).
119. Romero-Romero, S. et al. The Stability Landscape of de novo TIM Barrels Explored by a Modular Design Approach. *J. Mol. Biol.* **433**, 167153 (2021).
120. Doerr, S., Harvey, M. J., Noé, F. & De Fabritiis, G. HTMD: High-throughput molecular dynamics for molecular discovery. *J. Chem. Theory Comput.* **12**, 1845–1852 (2016).
121. Harvey, M. J., Giupponi, G. & Fabritiis, G. De. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* **5**, 1632–1639 (2009).
122. Maier, J. A. et al. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
123. Doerr, S. & De Fabritiis, G. On-the-Fly learning and sampling of ligand binding by high-throughput molecular simulations. *J. Chem. Theory Comput.* **10**, 2064–2069 (2014).
124. Harvey, M. J. & De Fabritiis, G. An implementation of the smooth particle mesh Ewald method on GPU hardware. *J. Chem. Theory Comput.* **5**, 2371–2377 (2009).
125. Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139**, 15102 (2013).
126. Buitinck, L. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
127. Weber, M. & Kube, S. Robust Perron cluster analysis for various applications in computational life science. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (eds. R. Berthold, M., Glen, R. C., Diederichs, K., Kohlbacher, O. & Fischer, I.) **3695 LNBI**, 57–66 (Springer Berlin Heidelberg, 2005).
128. Buch, I., Giorgino, T. & De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl Acad. Sci. USA.* **108**, 10184–10189 (2011).

Acknowledgements

Part of the work was funded by the Newton International Alumni Funding (2018-2023) awarded to A.A.J jointly by the Royal Society, the British Academy, and the Academy of Medical Sciences. J.D.M was partly funded by the French National Agency (CARTEGRIN ANR21-CE19-0017). S.R.R. was funded by a fellowship from the Alexander von Humboldt and Bayer Science & Education Foundations, NF was funded by the ERC Consolidator Grant 647548 and SD was funded by VolkswagenStiftung Grant 94747. The authors would like to thank Prof Birte Höcker for access to the DSC instruments. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High-Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b114cb (UID 210235). NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. We also thank Prof Thomas Scheibel for laboratory resources for executing part of this project. The authors specially thank Prof Richard Farndale for insightful discussion and help with the peptide synthesis.

Author contributions

A.A.J. and J.D.M. conceived the project and developed the experimental proposal. A.A.J. performed sequence analysis, made deductions based on these, made the figures, performed CD experiments, and also wrote the manuscript, S.R.R. performed D.S.C. experiments and the related data analysis, N.F. performed the MD simulations and analyzed the data,

S.D. clustered the collagen sequences using CD-HIT, V.S. performed the CD kinetic unfolding experiments of the peptides.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-54046-y.

Correspondence and requests for materials should be addressed to Abhishek A. Jalan.

Peer review information *Nature Communications* thanks Debora Monego, George Pantelopulos, Fei Xu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.