Briefings in Bioinformatics, 2025, **26(1)**, bbae622
https://doi.org/10.1093/bib/bbae622
Problem Solving Protocol

Diffusion model assisted designing self-assembling collagen mimetic peptides as biocompatible materials

Xinglong Wang[1,2], Kangjie Xu[1,2], Lingling Ma[1,2], Ruoxi Sun[1,2], Kun Wang[1,2], Ruiyan Wang[3], Junli Zhang[3], Wenwen Tao[3], Kai Linghu[1,2], Shuyao Yu[1,2], Jingwen Zhou[1,2,4,*]

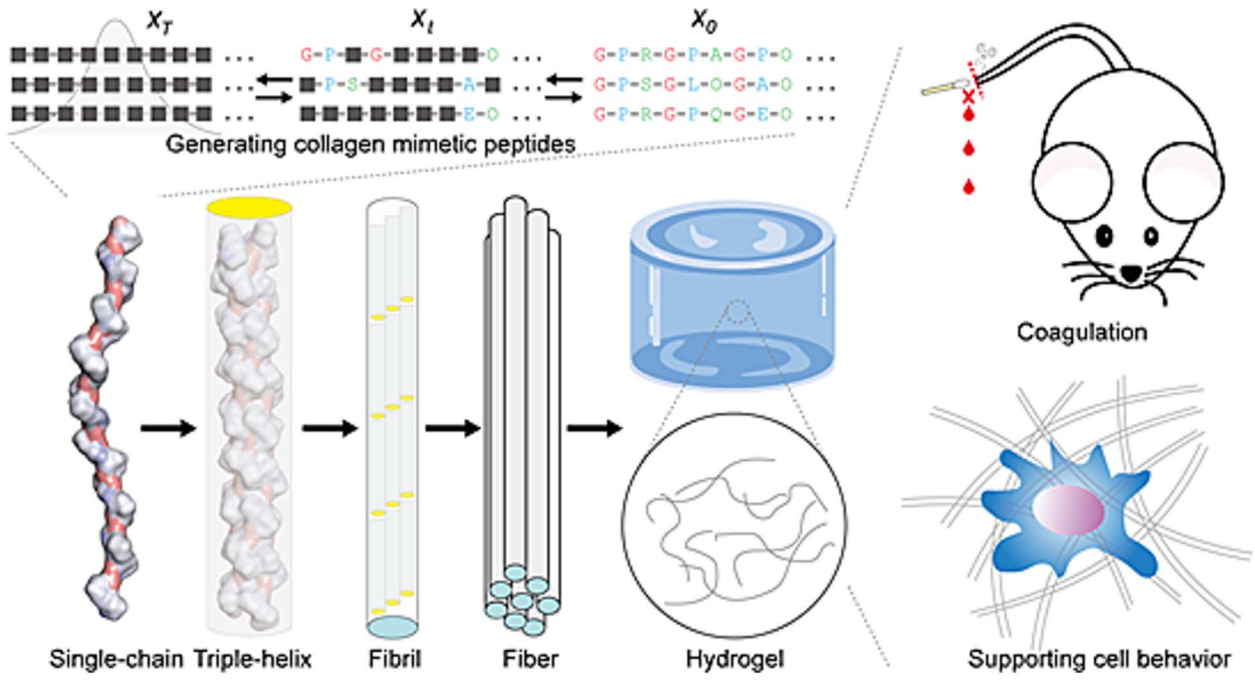
[1]Engineering Research Center of Ministry of Education on Food Synthetic Biotechnology and School of Biotechnology, Jiangnan University, 1800 Lihu Road, Wuxi, Jiangsu 214222, China
[2]Science Center for Future Foods, Jiangnan University, 1800 Lihu Road, Wuxi, Jiangsu 214222, China
[3]Bloomage Biotechnology Corporation Limited, 678 Tianchen Road, Jinan, Shandong 250104, China
[4]Jiangsu Province Engineering Research Center of Food Synthetic Biotechnology, Jiangnan University, 1800 Lihu Road, Wuxi, Jiangsu 214222, China

*Corresponding author. Science Center for Future Foods, Jiangnan University, 1800 Lihu Road, Wuxi, Jiangsu 214222, China. E-mail: zhoujw1982@jiangnan.edu.cn

Abstract

Collagen self-assembly supports its mechanical function, but controlling collagen mimetic peptides (CMPs) to self-assemble into higher-order oligomers with numerous functions remains challenging due to the vast potential amino acid sequence space. Herein, we developed a diffusion model to learn features from different types of human collagens and generate CMPs; obtaining 66% of synthetic CMPs could self-assemble into triple helices. Triple-helical and untwisting states were probed by melting temperature (T_m); hence, we developed a model to predict collagen T_m, achieving a state-of-art Pearson's correlation (PC) of 0.95 by cross-validation and a PC of 0.8 for predicting T_m values of synthetic CMPs. Our chemically synthesized short CMPs and recombinantly expressed long CMPs could self-assemble, with the lowest requirement for hydrogel formation at a concentration of 0.08% (w/v). Five CMPs could promote osteoblast differentiation. Our results demonstrated the potential for using computer-aided methods to design functional self-assembling CMPs.

Graphical Abstract



Keywords: diffusion model; deep learning; collagen mimetic peptides; osteoblast differentiation; self-assembling CMPs

Received: March 28, 2024. **Revised:** July 29, 2024. **Accepted:** November 14, 2024

Introduction

Collagens are widely distributed in bones, muscles, and blood [1], containing 28 subtypes, most of which are type I, II, and III [2]. Collagens are important biomaterials for cosmetics [3], drug delivery [4], nutraceuticals [5, 6] and tissue engineering [7, 8]. The structural integrity of collagen is important for supporting its mechanical function and stability [9]. Collagen consists of Gly-X-Y repeats, the basic unit for the triple helix, where X and Y are usually proline and hydroxyproline [1]. Collagen assembles into a triple helical structure as tropocollagen, which fundamentally supports the formation of collagen fibres. Collagen fibres, in turn, support the mechanical function under physiological conditions and endow collagen with properties that make it ideal as a biocompatible material for biomedical applications [10]. Native collagens have inherent limitations including poor stability and immunogenicity, low mechanical strength, and weak bioactivity [11–13], addressing the need for designing functional collagen mimetic peptides (CMPs) to overcome these limitations [14].

The formation of collagen triple helices facilitates higher-order oligomer assembly [15]. The process is triggered by noncovalent interactions between molecules, leading to a stable supramolecular structure [16]. Collagen can form hydrogels to support cell proliferation and adhesion [17–21]. However, designing CMPs to mimic native collagens to self-assemble into triple helices remains challenging. The G-X-Y repeat is the basic unit supporting triple-helix formation, and the content of charged residues and the degree of proline hydroxylation are believed to determine collagen self-assembly [10, 22, 23]. Additionally, various factors, including sequential composition and incubation conditions, can influence collagen self-assembly [24]. Given that self-assembly is crucial for collagen's biological functions, designing self-assembled CMPs is of significant importance.

The generative Artificial Intelligence (AI)-based collagen design was first conducted by Khare *et al.*, who combined generative and supervised models for collagen sequences at with desired melting temperature (T_m) [14]. The transition between triple helix and untwisted states of collagen can be represented by the T_m value [25], highlighting the significance for accurate prediction of collagen T_m. The predictive accuracy of collagen T_m values was further improved by a Bidirectional Encoder Representation from Transformers (BERT)-based deep learning model, achieving an R^2 of 0.84 [26]. Besides AI-based design, amino acid sequence analysis and molecular dynamics (MD) simulation have been employed to investigate the mechanism underpinning collagen self-assembly [27, 28]. MD simulation–based studies highlight the role of proline hydroxylation in internal noncovalent interactions [22] and provide insight into mutation-related CMP design. However, MD simulation–based methods require extensive computational resources that limit their utility for exploring novel CMPs.

To address the difficulties in designing CMPs and to control their self-assembly behaviour at different temperatures, we developed a combinatorial AI-based method involving a generative model and a supervised model. First, we developed a diffusion model–based network [29] to identify the features of different types of human collagens and generate CMPs, and their T_m values were predicted using the supervised model, which learned amino acid sequence features from a collagen T_m dataset [30]. The model proved accurate and was implemented for the selection of CMPs based on their predicted T_m values. We tested *Escherichia coli* and *Pichia pastrois* as possible platforms for recombinant expression of CMPs and found that several CMPs could be directly expressed in both hosts. Additionally, we investigated the contribution of proline hydroxylation to triple helix formation. To identify samples that outperformed currently available collagen products, we tested the capability of CMPs to inducing osteoblastic differentiation and cell adhesion.

Results

Generating collagen mimetic peptides by combining generative and supervised models

Collagen consists of G-X-Y repeats that drive triple-helix formation [31]. We developed a generative model (ColDiff) based on a diffusion model to learn amino acid sequence features from natural collagens and generate functional CMPs. A 30 aa region of human collagen could assemble into a triple helix that was verified by crystallography [32, 33]. Meanwhile, most collagen samples in the collagen T_m dataset [30] varied in length from 27 to 33 aa (Supplementary Data 1). Therefore, we decided to focus on generating collagens 30 aa in length because this length is sufficient for triple-helix assembly, and T_m can be easily predicted. Our aim was to generate CMPs with diverse functions by exploring the sequential space of human collagens. We collected 28 types of human collagens and conducted fragmentation, resulting in continuous G-X-Y repeats that are 30 aa in length. This curated training set comprised 7270 sequences (Supplementary Data 1).

ColDiff is an unconditional diffusion model derived from our previously developed diffusion model [34]. Here, the sequences from the training set were extracted features using one-hot encoding (Fig. S1). The input samples were added with Gaussian noise in a series of steps, and the model learned both forward and reverse processes to recover samples in the original denoised state (Fig. 1A). The quality of the generated samples was evaluated by consulting the portion of continuous G-X-Y repeats within the generated sequences. The best model achieved by hyperparameter optimization was able to generate continuous G-X-Y repeat sequences at a proportion > 92% after 50 epochs of training (Fig. 1B and Fig. S2). The proportion of G-X-Y repeat sequences fluctuated between 99.4% and 99.8% within 1000–2000 training epochs (Fig. 1B), indicating a stable CMP-generating ability for ColDiff.

A total of 25 600 sequences were generated by ColDiff by learning features from fragments of 28 types of human collagens, using training parameters set at 2000 epochs. Multiple sequence alignment (MSA) was conducted to evaluate the similarity between the generated and training sequences. The resulting sequence logo revealed continuous G-P-O repeats as the most prominent motifs, with charged residues appearing as secondary logos in both the training and generated sequences (Fig. 1C and Fig. S3). Sequence diversity was analysed using principal components analysis (PCA), revealing two main clusters that encompassed the majority of both training and generated sequences (Fig. 1D). Only a few generated sequences (three sequences) were distinctly separate from the training sequences, indicating that >99.99% of the generated samples closely resembled the training samples. Analysis of different G-X-Y frequencies showed that the G-X-Y distribution between the training and generated samples achieved an r^2 value of 0.98 (Fig. 1E). These results suggest that the composition of the generated samples closely matches that of the training samples. Functional motifs such as GAOGEN, GLKGEN, and GLOGEN were found in the generated sequences at a frequency of 0.6% (Supplementary Data 1) [35], which is slightly lower than the 2.2% observed in the training set (Fig. 1F). These results suggested that the generated samples are functionally related to the training samples. Overall, the generated CMPs closely resembled the collagen fragments in the training set.

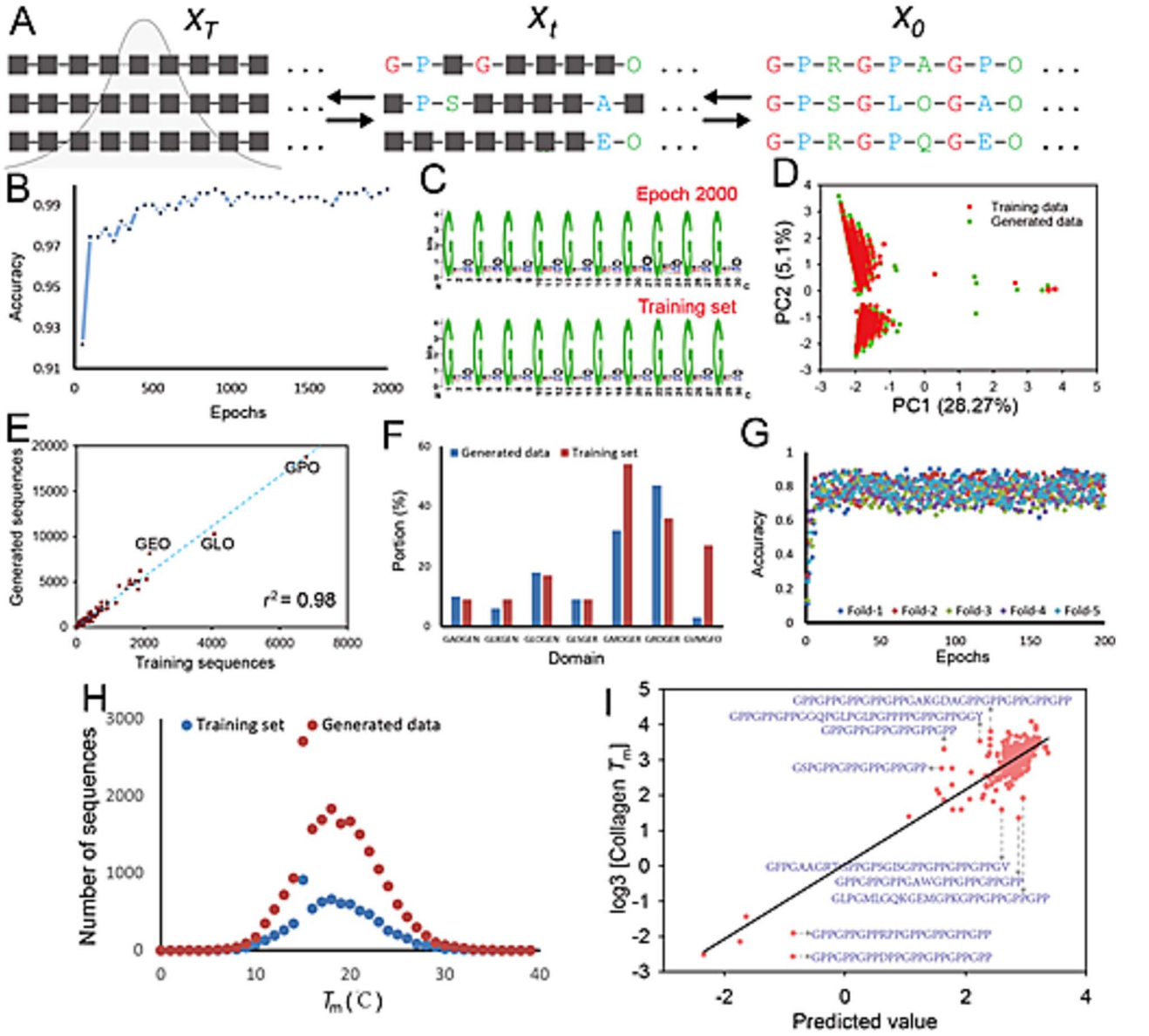


Figure 1. Developing a diffusion model for designing self-assembling CMPs. (A) Diffusion model development based on amino acid sequence learning. Sequence features of collagens were extracted by one-hot and by gradually adding Gaussian noise, the diffusion model learned the noising and denoising process to recover original sequences based on input sequences. In this case, the diffusion model was used to learn features from collagens and generate CMPs. (B) Accuracy of the diffusion model evaluated by calculating the proportion of continuous G-X-Y repeats from the generated sequences. (C) Sequence logos of generated sequences after training for 2000 epochs, and the original collagen dataset. Sequence logos were created using WebLogo (http://weblogo.berkeley.edu/logo.cgi). (D) Cluster analysis of sequences generated using ColDiff (25 600 samples) and from the training set (7270 samples). Cluster analysis conducted using PCA and KMeans cluster module. (E) Different G-X-Y frequencies of sequences generated by ColDiff and the training set, the correlation was analysed. (F) Functional domain distribution in training and generated sets. (G) Accuracy based on 5-fold cross-validation of ColNet. Accuracy was evaluated by calculating Pearson's correlation between predicted T_m and real T_m values. (H) Gaussian distribution of T_m values in training and generated sets. (I) Correlation between predicted T_m and real T_m values in the validation set. Sequences with highly distinct predicted and real values are shown.

The T_m of collagen is associated with triple-helix untwisting [36]. A supervised model ColNet was developed by training with a collagen T_m dataset for T_m prediction (Supplementary Data 2) [30]. T_m prediction was treated as a regression prediction task, and we integrated CNN, ResNet [37], and self-attention [38] for network architecture (Fig. S4). By conducting 5-fold cross-validation according to a previous study [26], our rebuilt network integrating ResNet and self-attention (SA) achieved a Pearson Correlation Coefficient (PCC) of 95% (r^2 = 90.2%) for predicted and real T_m values (Fig. 1G and Fig. S5). We noticed that ColNet failed with predicting few single mutation sequences, while the other sequences with distinct predicted and real T_m values did not exhibit obvious features (Fig. 1H and Fig. S6). The generated CMPs and collagens from the training set shared a similar Gaussian distribution for melting temperature, with the highest points of the two related curves being 23.5°C and 23.8°C, respectively (Fig. 1I), confirming the high correspondence between generated CMPs and native collagens.

Characterizing the self-assembly capacity of designed collagen mimetic peptides

The formation of the triple helix can be inferred from the Circular dichroism (CD) curves, which display a positive peak at 220–222 nm [39]. The transition temperature (T_m) of CMPs reflects

the triple-helix formation and untwisting state. Based on the T_m and functional domains of ColDiff-generated sequences, we selected 50 sequences with T_m values ranging from 3°C to 37°C for experimental validation, of which 30 had reported functional domains (Table 1) [35]. The synthetic peptides were dissolved in Phosphate Buffered Saline (PBS) (pH 7.5) at a final concentration of 0.2 mg/ml and incubated at 4°C for 12 h to enable their self-assembly. The CD spectra measured at 4°C showed that 33 samples displayed a positive peak at 220–222 nm, indicating a 66% success rate for ColDiff. The T_m values of the 33 samples were measured by heating them from 0°C to 70°C and recording the change of ellipticity at 220 nm. Most samples displayed T_m values varied from 7.5°C to 38.9°C (Fig. S7). The predicted and real T_m values achieved a PCC of 0.8 (Fig. 2A), indicating that the combined methods generate self-assembled CMPs. Since 25°C and 37°C are most relevant for bioengineering applications [14], we tested triple-helix formation for 15 samples with T_m values >25°C, and three samples with T_m values >37°C. Samples were pre-incubated at 25°C or 37°C for 12 h before CD spectroscopy. Our results indicated that the T_m values >5°C higher than the incubation temperature could stably sustain triple-helix formation (Fig. 2B), confirming a temperature-dependent triple helix to disordered state transition [36].

The above results indicated that not all continuous G-X-Y repeats were able to assemble into triple helices [14, 40]. Collagen self-assembly was associated with charged residues [16], the proportion of hydroxyproline [22], and GPO content [15]. The average number of charged residues within the 33 samples that formed triple helices was 6.06, compared to 5.76 for the 17 samples that did not form triple helices (Fig. 2C). Each triple helix–forming sequence contained ~4.51 hydroxyproline residues and 0.91 GPO motifs, whereas these values were 3.88 and 1, respectively, for sequences that did not form triple helices (Fig. 2C). Different G-X-Y portion analyses indicated a PCC of 0.81 for sequences that were able and not able to form triple helices (Fig. 2D). The top three G-X-Y motifs for sequences that formed triple helices were GPO, GLO, and GEO, while for those that did not form triple helices, they were GPO, GER, and GSO (Fig. 2D). Taken together, the proportion of hydroxyproline and specific G-X-Y content were important factors determining the self-assembly of CMPs, in addition to the basic continuous G-X-Y repeats.

Because all continuous G-X-Y sequences predicted using ColNet had positive T_m values, we aimed to explore the factors that determined the positive/negative values of the generated sequences. We noticed that sequences with negative values resulted from the disruption of G-X-Y repeats. To study this further, we used an initial sequence $(GPO)_3EPO(GPO)_4$ with a negative T_m value of −16°C to generate 8000 sequences, all predicted to have negative values (Supplementary Data 3). To explore the impact of disrupting G-X-Y repeats, we referred to the crystal structure of the collagen sequence $(Pro\text{-}Pro\text{-}Gly)_{10}$ [32] and disrupted this sequence by mutating Gly9 to Glu. MD simulation was conducted for 300 ns at 300 K; the RMSD > 5 nm took a portion of 2.89% for $(Pro\text{-}Pro\text{-}Gly)_{10}$, compared with 3.55% in the disrupted G-X-Y sequence (Fig. 2E). The average RMSD for $(Pro\text{-}Pro\text{-}Gly)_{10}$ and G9E were 0.427 and 0.439 nm, respectively. RMSF analysis indicated that the N- and C-termini and residues 9–20 within each chain displayed higher flexibility than other regions (Fig. 2F). Cluster analysis based on RMSD revealed a three-state transition: triple-helix twisting, one-chain untwisting, and completely untwisting (Fig. 2G). The partial and completely untwisting states occurred in 4.66% and 0.01% for $(Pro\text{-}Pro\text{-}Gly)_{10}$, respectively, compared to 6.98% and 0.45% for the G9E. These results were further validated by CD, which showed a negative peak at 220 nm following the disruption of G-X-Y repeat (Fig. 2F).

Recombinant expression of collagen mimetic peptides

Our initial goal was to identify ready-to-use biomaterials and develop a method for large-scale production. To this end, we first attempted to recombinant express CMPs in *E. coli*, by repeating the 30 aa segment up to 16 times to benefit the recombinant expression [41]. The results showed that 7 of the 50 recombinant CMPs could be expressed in soluble form (Fig. 3A and Fig. S8). Collagen solutions can be heated to 85°C followed by cold incubation [42], suggesting that they can recover their native structure after high temperature–induced untwisting. Therefore, we performed protease cleavage to remove the His-tag after purification, followed by heating to 85°C, cooling, centrifugation, and dialysis to remove contaminant proteins and cleaved tags (Fig. 3B). The resulting CMPs were of high purity (Fig. 3B). The CMPs were soluble expressed at a level of 0.1–0.2 mg/ml with bovine serum albumin (BSA) as a reference (Fig. S9), while sample-E43 achieved the highest expression level of 0.3–0.4 mg/ml (Fig. 3A).

Only a small portion of the CMPs could be expressed in *E. coli*, suggesting a low efficiency of this expression platform. Intracellular expression of CMPs in *E. coli* may suffer from multiprotease degradation [43, 44]. We therefore employed *P. pastoris* for secretory expression of CMPs [45] because *P. pastoris* contains only one endogenous protease that can degrade heterologous proteins after secretion [46]. Given that long protein chains may be easily degraded, these 30 aa sequences were replicated four times during gene synthesis, and CMPs with T_m values >25°C were selected for secretory expression in *P. pastoris* since retaining triple-helical conformation at temperature > 25°C can benefit practical uses. Six of the 15 samples were secreted with expression levels of 0.1–0.2 mg/ml (Fig. 3C and Fig. S10).

CD was carried out on the expressed proteins to confirm triple-helix formation [39]. The samples including sample-E3, sample-E14, and sample-E43 purified from *E. coli* and sample-P36 and sample-P43 purified from *P. pastoris* displayed positive peaks at 220–222 nm (Fig. S11). Compared with synthetic peptides, recombinantly expressed CMPs are much longer and lack proline hydroxylation [47]. These CMPs seemed hard to assemble into triple helix. To validate the impact of proline hydroxylation, the prolyl 4-hydroxylase (P4H) from *Bacillus anthracis* [48] was expressed and purified in *E. coli* (Fig. 3D) and used for *in vitro* hydroxylation of recombinantly expressed CMPs. The samples including sample-P1, sample-P6, and sample-P9 purified from *P. pastoris* formed triple helices following hydroxylation (Fig. 3E). Proline residues in collagen-like proteins were hydroxylated at a rate of 32.1%–58.3% (Fig. S12). Thus, proline hydroxylation may provide an alternative route for producing self-assembled CMPs, but the effort involved is considerable.

Fibre morphology of collagen mimetic peptides

Collagen higher-order assembly is important for developing biomaterials [15, 42]. Triple helix formation initially supported nanofiber formation and subsequently supported gel formation (Fig. 4A) [49]. Here, 38 samples able to form triple helices (excluding samples requiring P4H treatment) were subjected to gelation tests. Samples were solubilized in PBS at a concentration of 1 mg/ml and incubated at 4°C (Fig. S13). Sample-3, sample-E3, sample-32, and sample-E43 were able to form nanofibres after 12 h of incubation detected by transmission electron microscopy (TEM) at 100–200 nm, and porosity can be found from the CMP

Table 1. The sequence and properties of selected CMPs for validation in this studySequence.

	Predicted Tm	Real Tm	Domain	Total charge	Delta charge
GLOGPEGPRGIOGAOGENGIOGSKGEKGEO	25.93	29.67	GAOGEN	7.00	−1.00
GESGROGAEGAOGENGQOGPOGQRGPTGEQ	25.32	19.62	GAOGEN	6.00	−2.00
GVKGYRGAOGENGEDGLQGFOGLKGEOGIQ	15.85	17.03	GAOGEN	7.00	−1.00
GPOGTAGAOGENGSOGLOGESGPKGQRGFO	14.78	None	GAOGEN	4.00	0.00
GGOGRIGPRGAAGPOGLKGENGETGPOGPV	23.85	None	GLKGEN	5.00	1.00
GDRGAKGFOGLAGVSGPOGLKGENGMOGQM	22.29	38.90	GLKGEN	5.00	1.00
GVOGLQGLOGQOGLKGENGSVGFOGDKGEN	16.72	14.36	GLKGEN	5.00	−1.00
GMRGMOGFRGLDGDOGVOGLOGLKGENGSO	9.84	11.76	GLKGEN	6.00	0.00
GLOGENGVRGDOGPRGPOGFOGERGKOGPS	25.70	30.92	GLOGEN	7.00	1.00
GLOGENGDOGPRGHOGEDGEOGEKGRDGEO	23.86	19.36	GLOGEN	12.00	−4.00
GATGPOGPRGFKGPOGLOGENGATGEQGFO	13.23	23.90	GLOGEN	4.00	0.00
GLOGENGMQGLTGDRGPOGPOGPKGROGDF	10.03	13.42	GLOGEN	6.00	0.00
GPIGPIGPRGPOGLSGERGEOGTOGPTGPO	29.61	None	GLSGER	4.00	0.00
GEKGEQGEKGPRGLSGERGSRGVOGPLGQO	22.88	None	GLSGER	9.00	1.00
GDKGEIGEKGLOGLSGERGDIGNIGARGPO	17.05	None	GLSGER	9.00	−1.00
GVOGITGIRGHKGELGLOGLSGERGROGRO	15.18	15.17	GLSGER	8.00	4.00
GADGARGMOGERGROGTOGSOGLOGIRGDR	26.67	27.86	GMOGER	8.00	2.00
GEVGMOGERGEOGAQGLOGGQGQOGPRGPK	25.93	19.44	GMOGER	6.00	0.00
GPRGKOGMOGERGESGFQGPKGFEGPOGGO	15.32	11.17	GMOGER	7.00	1.00
GLOGMOGERGPKGRLGSOGENGEKGGIGFO	13.21	23.41	GMOGER	7.00	1.00
GPRGFKGAOGPRGDOGROGERGEOGLDGEO	32.30	27.65	GROGER	10.00	0.00
GQOGROGERGLOGIOGAOGLRGQOGPOGLD	29.60	27.80	GROGER	5.00	1.00
GLOGMLGPLGIMGSOGROGROGERGLAGQR	28.32	15.82	GROGER	5.00	3.00
GFTGAOGAKGQRGKOGPLGPOGPOGROGER	27.19	None	GROGER	6.00	4.00
GDOGIRGAOGLOGROGERGLTGPNGDOGFD	14.28	None	GROGER	7.00	−1.00
GPOGPQGPRGETGEOGDRGPRGROGERGAT	13.60	None	GROGER	9.00	1.00
GAKGSKGEKGFDGILGDVGROGERGSEGFO	10.83	None	GROGER	10.00	0.00
GPAGGOGVMGFOGPLGEKGNRGVOGLOGDQ	21.21	15.56	GVMGFO	4.00	0.00
GPKGDKGDOGPOGVMGFOGPKGEKGTQGSO	19.14	None	GVMGFO	7.00	1.00
GKOGPOGLDGTVGVMGFOGGKGEOGISGAO	10.84	14.34	GVMGFO	4.00	0.00
GSOGSSGPEGPOGEOGLAGEOGPVGEDGEA	38.16	33.82	6.00	−6.00	
GPOGPOGSQGMOGPEGPOGEOGPOGPOGLO	37.61	25.47	2.00	−2.00	
GEAGAQGPOGPOGNOGPOGVOGVDGPQGSS	36.92	None	2.00	−2.00	
GPOGPHGPLGSOGLOGEDGIOGLOGPOGSD	36.51	34.08	4.00	−2.00	
GLOGPQGVRGEOGDOGROGEOGPQGKOGEK	36.40	40.60	8.00	0.00	
GPTGPQGERGPRGEOGPOGPOGPOGLOGGS	36.39	36.01	4.00	0.00	
GPTGAOGPAGPOGROGQOGTOGQDGSOGLO	36.17	33.11	2.00	0.00	
GPIGKVGPAGSRGNOGQOGEOGLAGVOGQR	36.15	35.38	4.00	2.00	
GPTGPRGPOGPOGERGEDGEOGPRGPOGLO	36.15	13.94	7.00	−1.00	
GPKGEOGLTGPOGEOGPOGQOGPOGLOGVO	35.88	13.73	3.00	−1.00	
GRAGPRGROGFDGMAGDDGKOGLOGFIGFF	6.36	26.92	7.00	1.00	
GSRGLOGLOGLDGIOGQOGPKGIOGFOGSO	6.22	12.21	3.00	1.00	
GNRGCDGVOGLDGKOGEOGAKGEAGRDGAK	6.20	17.00	10.00	0.00	
GQKGRKGPKGLDGAOGFMGVSGFOGNOGAR	6.16	7.54	6.00	4.00	
GGOGPOGASGLDGMOGDMGEMGPOGIQGAR	5.80	None	4.00	−2.00	
GDEGPOGVAGLDGSOGSOGPOGFSGPOGHO	5.54	None	4.00	−2.00	
GPNGQTGARGPKGASFKOGTKGTOGFYGLF	5.51	None	4.00	4.00	
GPSGPAGSOGLDGAKGKQGPQGFKGVVGSO	4.65	None	4.00	2.00	
GPGGPNGAOIKMGREGEPGPTGPQGEDGPO	3.04	None	6.00	−2.00	
DROGAQGKGAHOGHPGQFRHHGTNGEMCQA	3.04	None	9.00	5.00	

The functional domains were as previous report.

sponges prepared by lyophilization using a scanning electron microscope (SEM) (Fig. 4B and Fig. S14). Sample E3 comprised 16 replicates of sample-3, but its band thickness was 55 nm, thinner than that of sample-3 (74 nm). The bandwidth for sample-32 and sample-E43 were 22 and 56 nm, respectively (Fig. S15).

Sample-E3 and sample-E43 formed fragile hydrogels at a very low concentration of 0.8 mg/ml after 8 h of incubation at 4°C (Fig. 4C). These hydrogels transitioned to a liquid phase when the temperature exceeded 15°C. In contrast, sample-3 and sample-32 required a higher concentration of 2 mg/ml to form hydrogels (Fig. 4D). Rheological analysis was conducted to evaluate the mechanical properties and stability of these hydrogels. The storage modulus (G') and the loss modulus (G") at low deformation amplitude were measured to determine the elastic and viscous contributions to the hydrogels' viscoelasticity. At 0.8 mg/ml, Sample-E3 and Sample-E43 exhibited minor distinctions in G' and G", indicating that the formed hydrogels were relatively weak and easily reverted to the liquid phase (Fig. 4C). When the concentration was adjusted to 2 mg/ml, the storage moduli more than doubled, indicating that all four samples formed stable hydrogels (Fig. 4D). Further increasing the concentration to 5 mg/ml and incubating for up to 7 days resulted in 11 samples forming

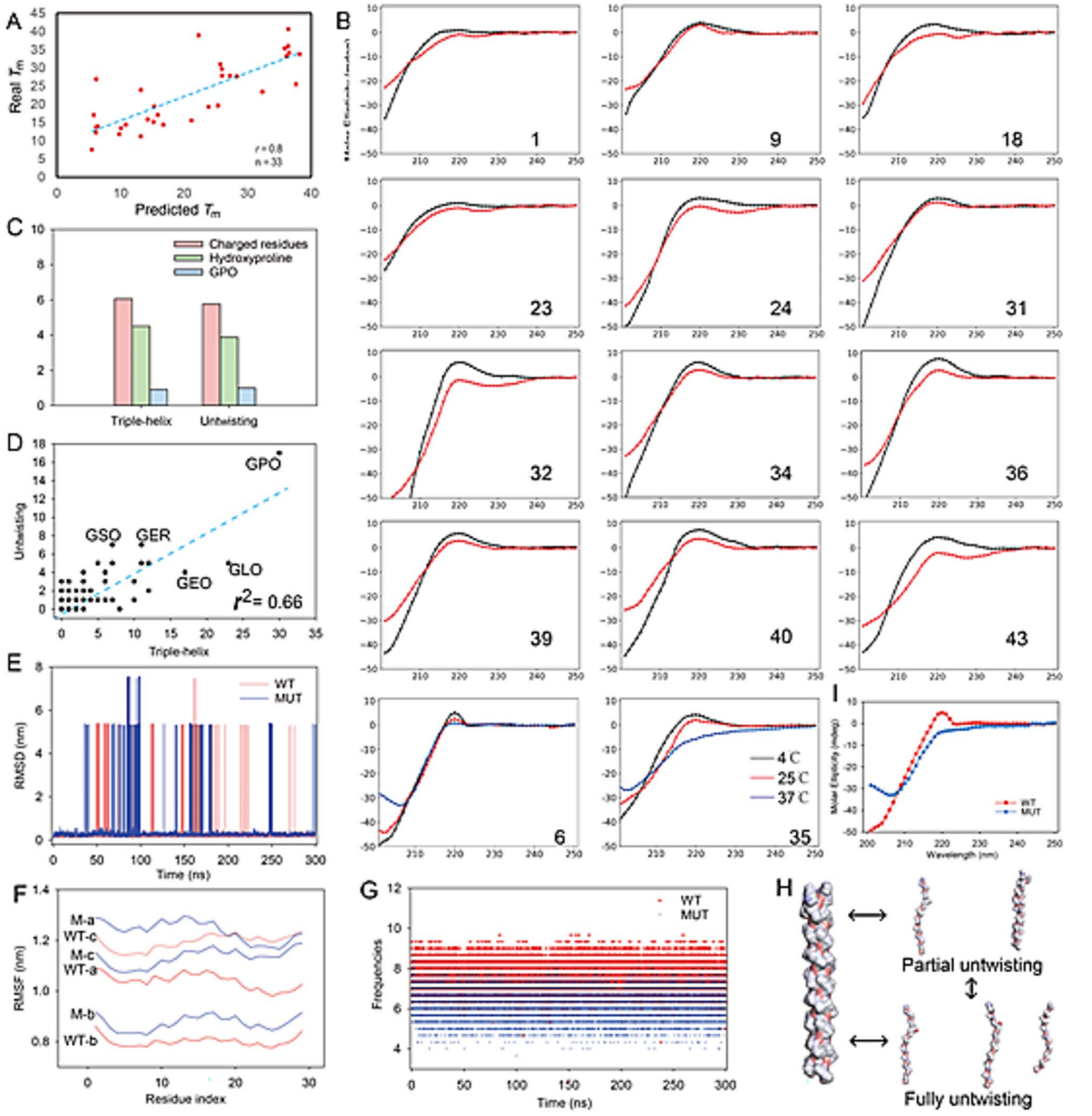


Figure 2. Triple-helical assembly of designed CMPs. (A) Correlations between predicted and real T_m values. (B) CD spectra of designed CMPs. Sample numbers are shown. (C) Sequence logos of the top 10 sequences with the highest predicted Tm values from randomly generated sequences. (D) The G-X-Y frequencies of sequences that were able or unable to form a triple helix, referred to as triple helix and untwisting, respectively. The RMSD (E), RMSF (F), and average hydrogen bond–forming frequencies (G) among the three independent chains of the given triple-helix obtained from MD simulation based on (Pro-Pro-Gly)10 (referred to as WT) and the G9E mutant (referred as MUT), respectively. The M- and WT-a, b, and c indicated the single chain of each complex in (F). MD simulation was carried out at 300 K using Gromacs-2020. (H) Analysis of triple-helix twisting and untwisting using MD simulation. Based on the trajectory, the representation of partial and full untwisting and their structures are shown. (I) CD spectra of (Pro-Pro-Gly)10 and the G9E mutant.

hydrogels (Fig. S15). However, these hydrogels remained fragile and could easily transition to the liquid phase upon shaking or temperature increase. Increasing the concentration to 50 mg/ml enhanced the hydrogels' rigidity and prevented them from converting to the liquid phase upon shaking. For samples that did not form hydrogels, large precipitates were observed (Fig. S16).

Collagens are believed to promote blood clotting by assisting platelet adhesion through binding to platelet receptors [42, 50] and subsequently inducing endogenous hemostatic mechanisms to aid wound healing [51]. We selected Sample-E3 and Sample-E43, which easily form hydrogels, for blood clotting tests and compared their performance with commercial rat tail collagen I. The sample sponges were incubated with mouse blood at a concentration of 2 mg/ml to assess their *in vitro* coagulation capability. After 30-min treatment at 25°C [50], Sample-E3 and Sample-E43 induced clotting in 96% and 92% of blood cells, respectively, compared to 81% for collagen I (Fig. 4E), suggesting the two samples with the capacity for inducing coagulation.

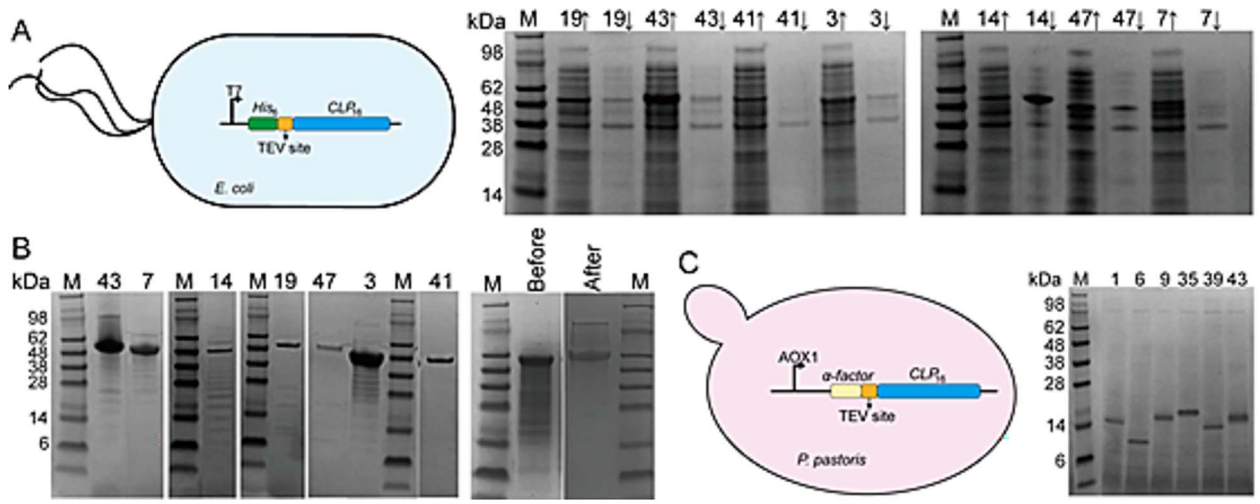


Figure 3. Recombinant expression of designed CMPs. (A) SDS-PAGE analysis of CMPs recombinantly expressed in *E. coli*. The expression cassette is shown, and the CMP number is included on the gel. (B) SDS-PAGE analysis of CMPs purified from *E. coli*. 'Before' and 'after' refer to sample 43 before and after heat treatment. (C) SDS-PAGE analysis of CMPs recombinantly expressed in *P. pastoris*.

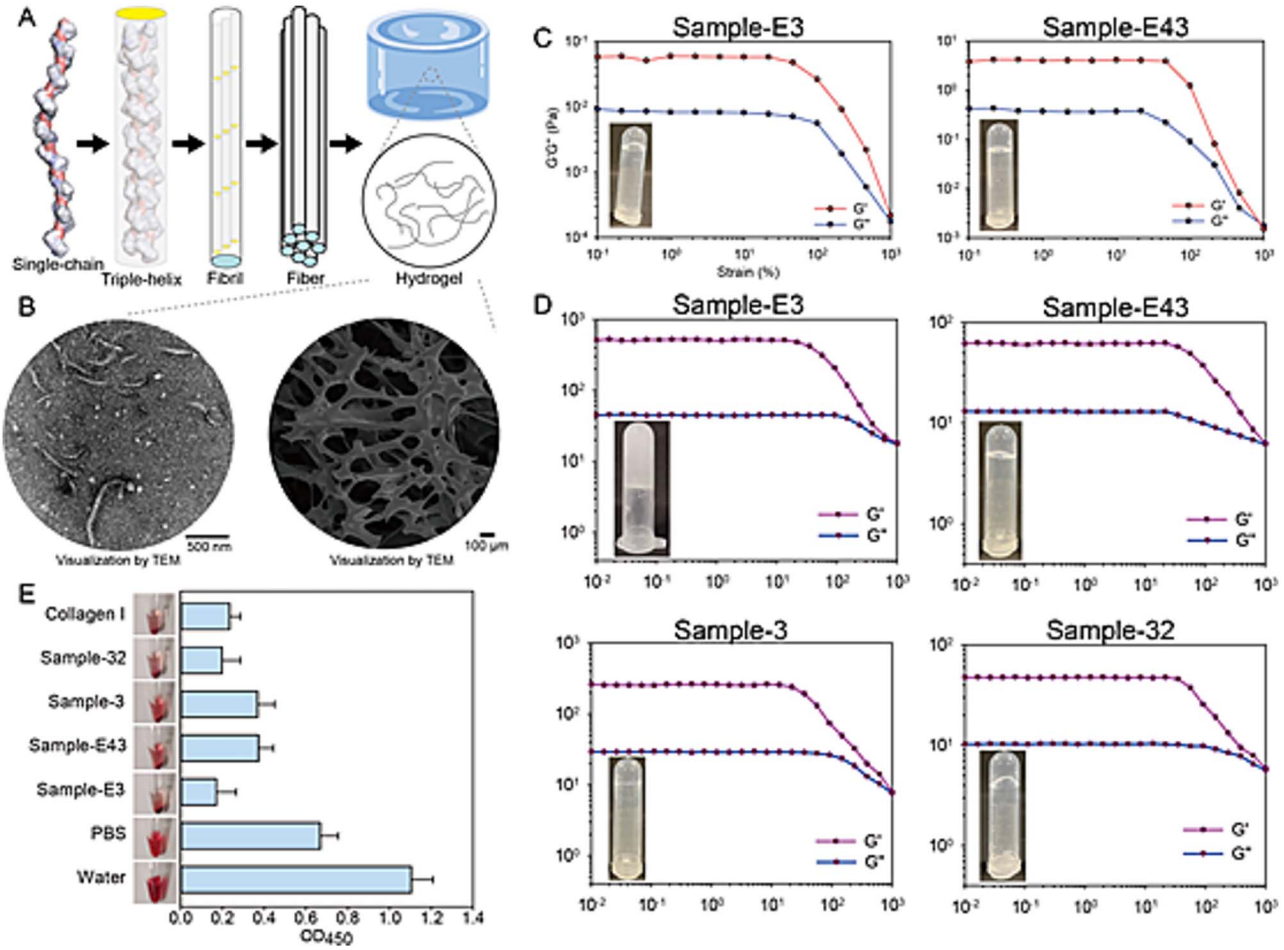


Figure 4. Morphology of designed CMPs. (A) Collagen assembly from triple-helices to higher-order oligomers. (B) TEM and images of samples E3 (left) and 3 (right). Samples were prepared at a concentration of 1 mg/ml. The samples were prepared as collagen sponges by lyophilization before SEM visualization. Visualization of the other samples by TEM and SEM were in Supplementary Fig. S14. Validation of hydrogel formation by oscillatory shear rheology analysis of CMPs at different concentrations, samples prepared at 0.8 mg/ml (C) and 2 mg/ml (D), respectively. (E) CMP sponge samples used for *in vitro* blood clotting by supplementing at 2 mg/ml, PBS (containing Ca2+) and water were used as a negative control, and rat tail collagen I was used as a positive control.

Assisting cell regeneration

Collagen assembles into a banded fibre that can support osteogenesis and matrix mineralization [15, 52, 53]. Nineteen samples were tested for their ability to assist osteoblastic differentiation, cell adhesion, and cell regeneration. The selected samples were prepared as sponges for MC-3 T3 cell cultivation. BSA and commercial collagen I served as negative and positive controls, respectively [15]. The designed CMPs and collagen I supplemented at 2 mg/ml promoted cell proliferation more effectively than BSA during the first 3 days of cultivation, and proliferation was similar by Day 5 (Fig. 5A), suggesting that the samples were not cytotoxic.

Cell adhesion analysis was carried out by adding CMPs and collagen I. The adhered cells were resuspended prior to cell counting. Collagen I and CMPs including sample-E3, sample-E14, sample-E43, sample-3, and sample-32 promoted >1-fold higher cell adhesion rate than BSA after 6-h cultivation (Fig. 5B). Genes *VCL* and *ACTN* are reportedly related to the formation of focal adhesion sites and cytoplasmic actin-binding protein occupation [15], and messenger RNA (mRNA) levels of *VCL* and *ACTN* can be used to evaluate osteoblastic differentiation. Sample-6, sample-9, sample-24, sample-E3, sample-E43, and collagen I exhibited ability to upregulate *VCL* and *ACTN*, the mRNA level of the given two genes were >1-fold higher than that using BSA (Fig. 5B). Meanwhile, the other 14 samples did not upregulate *VCL* or *ACTN* transcription relative to the BSA-negative control (Fig. 5B). In addition, the cell areas of CMP samples including sample-E3, sample-E14, sample-E43, sample-6, and type I collagen were >88% larger than whose when using BSA alone (Fig. 5C and D).

The recombinant expressed sample-E3 and sample-E43, which exhibited activity for inducing cell differentiation, were further assessed for their biocompatibility over a cultivation period of up to 7 days. Our results indicated that cell growth maintained a stable trend within 7 days of cultivation (Fig. S17A). The proliferation rates were ∼4-fold and 10% during the first 5 days and last 2 days, respectively. SDS-PAGE analysis revealed that both samples were stable with minor degradation at 25°C during 7 days' incubation, losing <20% of their original state by the seventh day (Fig. S17B).

Discussion

Designing self-assembling CMPs remains challenging, here, we developed a diffusion model to learn features from the retrieved segments of human collagen to generate diverse CMPs, and subsequently developed ColNet for collagen T_m prediction to select CMPs with desirable T_m values. CMPs recombinantly expressed in *E. coli* and *P. pastoris* pave the way for large-scale production. Additionally, we found that CMPs with up to 16 repeats can undergo triple helix formation even without proline hydroxylation. Nanofibre and hydrogel formation were investigated, and four samples could rapidly form hydrogels at low concentrations (≤ 2 mg/ml). In addition, we found the CMPs can induce osteoblastic differentiation mimic natural collagens at a comparable rate.

Generative models have been implemented to design protein-binding molecules and optimize functional areas [54, 55]. For example, a sequential-based generative network, proteinGAN, was used to design malate dehydrogenase (MDH) by learning from its homologues [56], while ProtGPT2 utilized an unsupervised language model to generate novel sequences that adhere to the principles of natural ones [57]. Specifically developed for designing collagens, ColGen-GA was built on a T_m prediction model to control the self-assembly of generated collagens [14]. Inspired by ColGen-GA, we optimized ColDiff for generating novel sequences, achieving a 66% self-assembly rate among the generated sequences. Additionally, we enhanced ColNet for predicting T_m values, achieving an average PCC of 95%, which surpasses the reported state-of-the-art model, CollagenTransformer (91.6%) [26]. We critically assessed the success rate of ColNet using 33 synthetic collagen-mimetic peptides (CMPs), finding a PCC of 0.8, which is lower than the PCC of 0.96 obtained in ColGen-GA based on five samples (CP1-CP5) [14]. A limitation of ColNet is that it predicts a negative T_m value if the G-X-Y repeats are disrupted in the input sample. This limitation could potentially be addressed by incorporating more diverse training data.

Collagen self-assembly and its functional motifs are fundamental components that support cell adhesion and regeneration [15, 35, 58]. Among the 50 selected CMPs, only five sequences were capable of supporting osteoblast differentiation. Four of these five sequences contained at least one known functional motif, such as GAOGEN, GLKGEN, GLOGEN, and GROGER [35], which can bind to cell surface receptors and activate behaviors like differentiation. Importantly, we showed that recombinant expressed sample-E3 and-E43, which replicated the sequence 16 times of sample-3 and sample-43, exhibited activity for inducing cell differentiation, suggesting that replicating the motifs may enhance the potential interactions between CMPs and cell surface receptors. Furthermore, four samples exhibited similar activity compared to commercial collagen I, while sample-9 and sample-24 showed higher activity. Collagens self-assembly into hydrogels is highly dependent on concentration, typically requiring concentrations exceeding 5 mg/ml [15, 17, 21, 42, 50, 58, 59]. Here, the CMPs with biofunctions were able to self-assemble into higher-order structures at lower concentrations, potentially supporting cell adhesion and stimulating differentiation through interactions involving their functional motifs. Moreover, these novel protein-based materials required lower concentrations to achieve their functions compared to previously reported elastin- and collagen-based materials, as well as polycaprolactone or chitosan-based materials [60–63]. These results indicate that the designed CMPs may become candidates supporting bone tissue engineering.

In this study, we adopted AI tools to explore *de novo* generation of functional, self-assembling CMPs at specific temperatures. The designed CMPs could assemble into nanofibres and hydrogels at low concentrations [17, 18], decreasing the dependence of their functions on dosage and ambient temperature. The designed CMPs were more effective at supporting cell regeneration than commercially available collagen I. Moreover, many CMPs retained the triple-helical conformation in the liquid phase. Liquid-phase collagens are in much demand for use as dietary supplements [64]. Further investigations are needed to explore the potential uses of the designed CMPs.

Materials and methods

Network architecture and model evaluation

The dataset for training the diffusion model was based on 30 aa G-X-Y repeats from natural human collagen. We collected 44 sequences from 28 types of human collagen to build the library, and the training set contained 7270 sequences (Supplementary Data 1). ColDiff is a diffusion model [29] integrating UNet for feature extraction on the encoder side and sample size recovery on the decoder side. The features of collagen sequences were extracted using the one-hot encoding method. We attempted CNN, ResNet, and SA for network architecture. The initial data point was denoted as X_0, and variable X_t was achieved

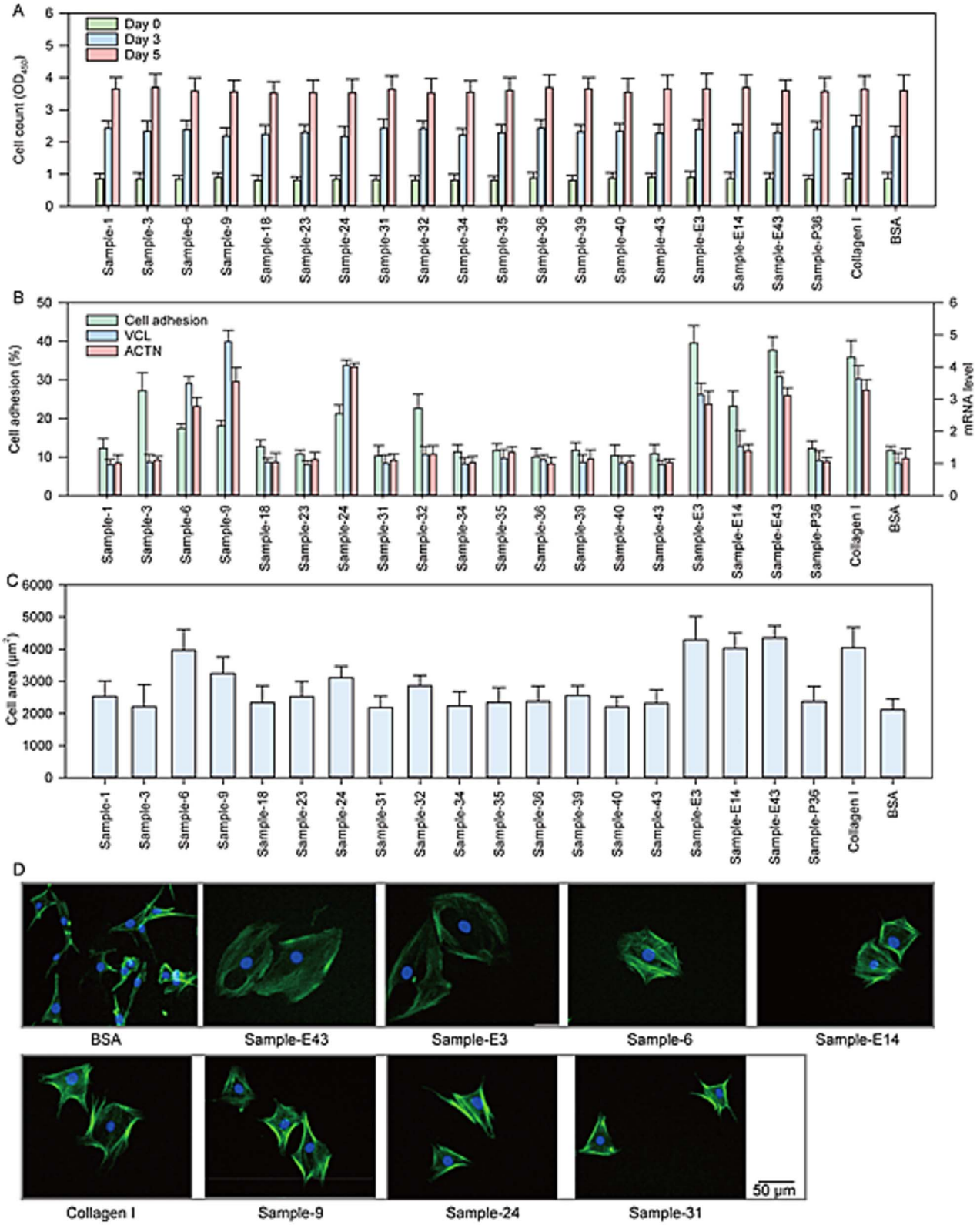


Figure 5. Cell regeneration ability of designed CMPs. (A) Cell proliferation rate determined by cultivating MC3T3-E1 cells supplemented with 2 mg/ml samples and cell counting. (B) Cell adhesion assay and correlated mRNA level measured by cultivating MC3T3-E1 cells supplemented with CMPs or collagen I hydrogels. (C) Cell area calculated using ImageJ after cultivation. (D) Staining of MC3T3-E1 cells with phalloidin and visualization by fluorescence microscopy (scale bar = 50 μm).

through adding Gaussian noise to X_0 for T steps. By reversing the noise, the network was trained to recover the original data.

The formula for ResNet [37] is shown below, where x stands for the input, F(x) is the output from the layer, Wi is the parameter to feed the CNN layer, and Ws represents certain convolution configurations to make the dimensions of input and output identical.

$$Y = F\left(x, (Wi)\right) + Ws * x$$

The formula for the self-attention mechanism [38] is shown below, where Q, K, and V are vectors of queries, keys, and values of dimension d_k, where d_k is the size of the attention key.

$$Attention\,(Q, K, V) = softmax\left(\frac{QKT}{\sqrt{dk}}\right)V.$$

The ColNet network for collagen T_m prediction was integrated by ResNet and self-attention through network optimization (Fig. S4). ColNet learned collagen features [30] and predicted T_m values by treating the process as a regression prediction problem. The learning rate and batch size were optimized based on the mean squared error between predicted and real T_m values. The accuracy of the model was defined as Pearson's correlation between predicted and real T_m values.

Collagen mimetic peptide synthesis and recombinant expression and purification

CMPs were chemically synthesized by Genscript (Nanjing, China) and Sangon (Shanghai, China) (Supplementary Data 4). Recombinant expression of CMPs was carried out in *E. coli* BL21 (DE3) using pET-22b (+) and *P. pastoris* GS115 using pPIC9k. All genes encoding CMPs were optimized and synthesized by Genscript (Nanjing, China) and cloned into pET-22b(+) via *NdeI-BlpI* sites and into pPIC9k via *SnaBI-AgeI* sites. Plasmids were transformed into *E. coli* using chemical transformation, and cultivated using Terrific Broth supplemented with 50 μg/ml ampicillin, followed by IPTG induction when the OD_{600} value reached 1.0. The cultivation was continued at 20°C for 20 h. Plasmids were transformed into *P. pastoris* using electroporation [65]. Minimal dextrose medium (20 g/L glucose, 13.4 g/L YNB, 4×10^{-5} g/L biotin and 20 g/L agar) was used for selecting histidine-defective strains. Yeast strains were cultivated using yeast extract peptone dextrose medium (1% yeast extract, 2% peptone, 2% glucose). Protein purification was carried out using affinity chromatography with a His-Trap column (GE Healthcare, New York, USA) and subjected to gel filtration using a Superdex 75 column (GE Healthcare) and eluted with PBS (pH 7.5).

CD spectroscopy

The sample solution was prepared at 0.2 mg/ml in PBS (pH 7.5) and incubated at 4 °C for 12 h before CD measurement. CD spectra were recorded using a Photophysics Chirascan instrument with a Peltier temperature controller (Model 110-OS; Hellma, Shanghai, China) at 4°C, and quartz cuvettes with an optical path length of 1 mm (Model 110-OS; Hellma). The wavelength for scanning ranged from 190 to 250 nm. The change in ellipticity at 220 nm during heating (0°C–70°C at 1°C/min) was recorded, and T_m values were calculated based on melting curves.

Transmission electron microscopy and scanning electron microscope

Samples were prepared from 1 to 100 mg/ml in PBS (pH 7.5) and incubated at 4°C. The obtained solution was loaded onto a copper grid and allowed to absorb for 1 min, and the excessive solvent was removed using filter paper. Samples were negatively stained using 0.75% phosphotungstic acid prior to imaging using a Hitachi H-7650 electron microscope (Hitachi, Tokyo, Japan). The sample width was evaluated using ImageJ (https://imagej.net/downloads). The morphology and microstructure of the CMP sponge were characterized using SEM SU8220 (Hitachi). The CMP gel was prepared to sponge by lyophilization. The dried samples were loaded to SEM pucks and proceeded to imaging.

Blood clotting assays

CMPs were used to prepare hydrogels for swelling rate determination [50]. Hydrogels were cultivated with mouse blood (Gibco, Shanghai, China) at 25°C and rinsed with PBS. Excess liquid was removed using filter paper, and the initial weight of hydrogels and after blood swelling was recorded to calculate the swelling rate. In addition, CMP sponges were mixed with anticoagulant mouse blood at a concentration of 2 mg/ml, and the mixture was supplemented with 10 mM $CaCl_2$. Coagulation testing was carried out at 37°C for 10 min. The absorbance at 540 nm was recorded before and after coagulation to evaluate the blood clotting speed. The swelling rate was calculated using the following formula:

$$Swellingrate(\%) = 1 - \frac{Weight\ of\ hydrogel\ before\ cultivation}{Weight\ of\ hydrogel\ after\ cultivation.}$$

Cell adhesion and proliferation assay

Cell adhesion and proliferation were assessed using MC-3 T3 cells [15]. A 24-well cultivation plate was filled with 10^4 MC-3 T3 cells (ATCC) and cultivated using Gibco MEM (Invitrogen, Shanghai, China), 10% (v/v) foetal bovine serum and 1% (v/v) penicillin–streptomycin supplemented with 2 mg/ml lyophilized dried CMP or collagen I. A CCK-8 kit was used to record the cell proliferation rate during 5 days of cultivation. For cell adhesion testing, CMP and collagen I samples were supplemented at 2 ml to cultivate cells. During cell adhesion testing, 10^4 precultivated MC-3 T3 cells were loaded onto the plate and cultivated for 12 h to allow cell adhesion, the plate was washed with PBS, cultivation medium was added, and cell counting was performed using a CCK-8 Kit. Gene expression levels were measured by real-time quantitative PCR (RT-qPCR). Primers were as described in a previous study [15] and are listed in Table S1. Gene expression levels were normalized against *GAPDH*. Cell adhesion and proliferation were calculated using the following formula:

$$Celladhesion(\%) = \frac{Adherent\ cells}{Seeding\ cells}$$

$$Proliferationrate(\%) = \frac{Total\ cells\ after\ cultivation}{Total\ cells\ before\ cultivation}$$

Molecular dynamics simulation

The crystal structure of $(Pro\text{-}Pro\text{-}Gly)_{10}$ was obtained from the Protein Data Bank (PDB; 1K6F). The structure with the G9E mutation was generated using Rosetta Remodel [66]. MD simulation was carried out using Gromacs-2020 [67, 68], which has been successfully utilized in the literature [69, 70]. The triple-helix-formed collagen was embedded with FF14sb force field [71], and the simulation box was filled with SPC/E water in a cubic box where the distance between the edge of the cubic box and the protein was 12 Å. The rigid bonds were used as the constraints for

water molecules. The system was neutralized by Na^+ and Cl^- ions, and the total number of atoms in the system was ~150 000. The system was initially energy-minimized using the steepest descent method. Hydrogen bond lengths were restrained using SETTLE and LINCS algorithms [72]. Long-range electrostatic interactions were calculated with the fourth-order particle mesh Ewald (PME) method [73], while nonbonded interactions had a cutoff distance of 1.2 nm, switching at 1 nm. The system was equilibrated by isochoric-isothermal ensemble (NVT) at 300 K for 100 ps and proceeded to isothermal-isovolumetric ensemble (NPT) at 300 K for 200 ps. The pressure $p = 1$ atm and a time step of 1 fs were used for the given two methods. The simulation was carried out at 300 K for 300 ns using a time step of 2 fs, and the trajectory was used for analysis. We calculated the RMSD, RMSF, and hydrogen bond formation frequencies at the atomic level by selecting the group of 'Protein'.

Key Points

- CMPs were designed by combinatory uses of the diffusion model and supervised model.
- Synthetic CMP peptides confirmed the designed CMPs can in high correlation to the predicted Tm values.
- The designed CMPs can self-assemble into higher order, forming nanofibre and hydrogel.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Funding

This study was funded by the Natural Science Foundation of Jiangsu Province (BK20202002), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (Grant No. SN-ZJU-SIAS-0013), China Postdoctoral Science Foundation (2023 M741403), Jiangsu Funding Program for Excellent Postdoctoral Talent (2023ZB037), the National First-class Discipline Program of Light Industry Technology and Engineering (QGJC20230102).

Data availability

All code and data used in this study can be found in the GitHub repository: https://github.com/wangxinglong1990/Collagen_design.

Supporting information

The data supporting the findings of this study are available within the article and its Supplementary Information. Other data and reagents are available from the corresponding authors upon reasonable request. Source data are provided in this paper.

References

1. Bielajew BJ, Hu JC, Athanasiou KA. Collagen: quantification, biomechanics and role of minor subtypes in cartilage. *Nat Rev Mater* 2020;**5**:730–47. https://doi.org/10.1038/s41578-020-0213-1.
2. Sorushanova A, Delgado LM, Wu Z. *et al.* The collagen Suprafamily: from biosynthesis to advanced biomaterial development. *Adv Mater* 2019;**31**:e1801651. https://doi.org/10.1002/adma.201801651.
3. Han S-B, Won B, Yang S-c. *et al. J Ind Eng Chem* 2021;**98**:289–97. https://doi.org/10.1016/j.jiec.2021.03.039.
4. An B, Lin Y-S, Brodsky B. Collagen interactions: drug design and delivery. *Adv Drug Deliv Rev Reviews* 2016;**97**:69–84. https://doi.org/10.1016/j.addr.2015.11.013.
5. Ahmad MI, Li Y, Pan J. *et al.* Collagen and gelatin: structure, properties, and applications in food industry. *Int J Biol Macromol* 2023;**254**:128037–51. https://doi.org/10.1016/j.ijbiomac.2023.128037.
6. Pogačnik T, Žmitek J, Hristov H. *et al.* The effect of a 12-week dietary intake of food supplements containing collagen and MSM on dermis density and other skin parameters: a double-blind, placebo-controlled, randomised four-way study comparing the efficacy of three test products. *J Funct Foods* 2023;**110**:105838–51. https://doi.org/10.1016/j.jff.2023.105838.
7. Wang Y, Wang Z, Dong Y. Collagen-based biomaterials for tissue engineering. *ACS Biomater Sci Eng* 2023;**9**:1132–50. https://doi.org/10.1021/acsbiomaterials.2c00730.
8. Chung H, Choi J-K, Hong C. *et al.* A micro-fragmented collagen gel as a cell-assembling platform for critical limb ischemia repair. *Bioact Mater* 2024;**34**:80–97. https://doi.org/10.1016/j.bioactmat.2023.12.008.
9. Kirkness MWH, Lehmann K, Forde NR. Mechanics and structural stability of the collagen triple helix. *Curr Opin Chem Biol* 2019;**53**:98–105. https://doi.org/10.1016/j.cbpa.2019.08.001.
10. Irastorza A, Zarandona I, Andonegi M. *et al.* The versatility of collagen and chitosan: from food to biomedical applications. *Food Hydrocoll* 2021;**116**:106633–44. https://doi.org/10.1016/j.foodhyd.2021.106633.
11. Fertala A. Three decades of research on recombinant collagens: reinventing the wheel or developing new biomedical products? *Bioengineering* 2020;**7**:155–81. https://doi.org/10.3390/bioengineering7040155.
12. Wang H. A review of the effects of collagen treatment in clinical studies. *Polymers* 2021;**13**:3868–88. https://doi.org/10.3390/polym13223868.
13. Saha S, Costa RC, Silva MC. *et al.* Collagen membrane functionalized with magnesium oxide via room-temperature atomic layer deposition promotes osteopromotive and antimicrobial properties. *Bioact Mater* 2023;**30**:46–61. https://doi.org/10.1016/j.bioactmat.2023.07.013.
14. Khare E, Yu C-H, Gonzalez Obeso C. *et al.* Discovering design principles of collagen molecular stability using a genetic algorithm, deep learning, and experimental validation. *Proc Natl Acad Sci USA* 2022;**119**:e2209524119–27. https://doi.org/10.1073/pnas.2209524119.
15. Hu J, Li J, Jiang J. *et al.* Design of synthetic collagens that assemble into supramolecular banded fibers as a functional biomaterial testbed. *Nat Commun* 2022;**13**:6761. https://doi.org/10.1038/s41467-022-34127-6.
16. Leo L, Bridelli MG, Polverini E. Insight on collagen self-assembly mechanisms by coupling molecular dynamics and UV spectroscopy techniques. *Biophys Chem* 2019;**253**:106224–37. https://doi.org/10.1016/j.bpc.2019.106224.
17. Bera S, Cazade P-A, Bhattacharya S. *et al.* Molecular engineering of rigid hydrogels co-assembled from collagenous helical peptides based on a single triplet motif. *ACS Appl Mater Interfaces* 2022;**14**:46827–40. https://doi.org/10.1021/acsami.2c09982.

18. Sarrigiannidis SO, Rey JM, Dobre O. *et al.* A tough act to follow: collagen hydrogel modifications to improve mechanical and growth factor loading capabilities. *Materials Today Bio* 2021; **10**:100098–120. https://doi.org/10.1016/j.mtbio.2021.100098.
19. Somaiah C, Kumar A, Mawrie D. *et al.* Collagen promotes higher adhesion, survival and proliferation of mesenchymal stem cells. *PLoS One* 2015;**10**:e0145068–83. https://doi.org/10.1371/journal.pone.0145068.
20. Pawelec KM, Best SM, Cameron RE. Collagen: a network for regenerative medicine. *J Mater Chem B* 2016;**4**:6484–96. https://doi.org/10.1039/C6TB00807K.
21. Qiu Y, Qiu S, Deng L. *et al.* Biomaterial 3D collagen I gel culture model: a novel approach to investigate tumorigenesis and dormancy of bladder cancer cells induced by tumor microenvironment. *Biomaterials* 2020;**256**:120217–30. https://doi.org/10.1016/j.biomaterials.2020.120217.
22. Rappu P, Salo Antti M, Myllyharju J. *et al.* Role of prolyl hydroxylation in the molecular interactions of collagens. *Essays Biochem* 2019;**63**:325–35. https://doi.org/10.1042/EBC20180053.
23. Hulmes DJS, Miller A, Parry DAD. *et al.* Analysis of the primary structure of collagen for the origins of molecular packing. *J Mol Biol* 1973;**79**:137–48. https://doi.org/10.1016/0022-2836(73)90275-1.
24. Yamamura N, Sudo R, Ikeda M. *et al.* Effects of the mechanical properties of collagen gel on the In vitro formation of microvessel networks by endothelial cells. *Tissue Eng* 2007;**13**:1443–53. https://doi.org/10.1089/ten.2006.0333.
25. Persikov AV, Ramshaw JAM, Kirkpatrick A. *et al.* Electrostatic interactions involving lysine make major contributions to collagen triple-helix stability. *Biochemistry* 2005;**44**:1414–22. https://doi.org/10.1021/bi048216r.
26. Khare E, Gonzalez-Obeso C, Kaplan DL. *et al.* CollagenTransformer: end-to-end transformer model to predict thermal stability of collagen triple helices using an NLP approach. *ACS Biomater Sci Eng* 2022;**8**:4301–10. https://doi.org/10.1021/acsbiomaterials.2c00737.
27. Strawn R, Chen F, Jeet Haven P. *et al.* To achieve self-assembled collagen mimetic fibrils using designed peptides. *Biopolymers* 2018;**109**:e23226–36. https://doi.org/10.1002/bip.23226.
28. Hafner AE, Gyori NG, Bench CA. *et al.* Modeling Fibrillogenesis of collagen-mimetic molecules. *Biophys J* 2020;**119**:1791–9. https://doi.org/10.1016/j.bpj.2020.09.013.
29. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *arXiv* 2020;**574**:6840–51. https://doi.org/10.1089/ten.2006.0333.
30. Yu C-H, Chen W, Chiang Y-H. *et al.* End-to-end deep learning model to predict and design secondary structure content of structural proteins. *ACS Biomater Sci Eng* 2022;**8**:1156–65. https://doi.org/10.1021/acsbiomaterials.1c01343.
31. Ramshaw JAM, Shah NK, Brodsky B. Gly-X-Y tripeptide frequencies in collagen: a context for host–guest triple-helical peptides. *J Struct Biol* 1998;**122**:86–91. https://doi.org/10.1006/jsbi.1998.3977.
32. Berisio R, Vitagliano L, Mazzarella L. *et al.* Crystal structure of the collagen triple helix model [(pro-pro-Gly)10]3. *Protein Sci* 2002;**11**:262–70. https://doi.org/10.1110/ps.32602.
33. Hua C, Zhu Y, Xu W. *et al.* Characterization by high-resolution crystal structure analysis of a triple-helix region of human collagen type III with potent cell adhesion activity. *Biochem Bioph Res Co* 2019;**508**:1018–23. https://doi.org/10.1016/j.bbrc.2018.12.018.
34. Wang X, Xu K, Tan Y. *et al.* Deep learning-assisted Design of Novel Promoters in *Escherichia coli*. *Adv Genet* 2023;**4**:2300184–95. https://doi.org/10.1002/adma.202412059.
35. Hamaia SW, Pugh N, Raynal N. *et al.* Mapping of potent and specific binding motifs, GLOGEN and GVOGEA, for integrin using collagen toolkits II and III. *J Biol Chem* 2012;**287**:26019–28. https://doi.org/10.1074/jbc.M112.353144.
36. Fujii KK, Taga Y, Takagi YK. *et al.* The thermal stability of the collagen triple helix is tuned according to the environmental temperature. *Int J Mol Sci* 2022;**23**:2040. https://doi.org/10.3390/ijms23042040.
37. He K, Zhang X, Ren S. *et al.* Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016;**2015**:770–8. https://doi.org/10.1006/jsbi.1998.3977.
38. Vaswani A, Shazeer NM, Parmar N. *et al.* Line: attention is all you need. *In: NeurIPS* 2017;6000–10.
39. Drzewiecki KE, Grisham DR, Parmar AS. *et al.* Circular dichroism spectroscopy of collagen Fibrillogenesis: a new use for an old technique. *Biophys J* 2016;**111**:2377–86. https://doi.org/10.1016/j.bpj.2016.10.023.
40. Hu J, Wang J, Zhu X. *et al.* Design strategies to tune the structural and mechanical properties of synthetic collagen hydrogels. *Biomacromolecules* 2021;**22**:3440–50. https://doi.org/10.1021/acs.biomac.1c00520.
41. Ma L, Liang X, Yu S. *et al.* Expression, characterization, and application potentiality evaluation of recombinant human-like collagen in *Pichia pastoris*. *Bioresour Bioprocess* 2022;**9**:119. https://doi.org/10.1186/s40643-022-00606-3.
42. Kumar VA, Taylor NL, Jalan AA. *et al.* A nanostructured synthetic collagen mimic for hemostasis. *Biomacromolecules* 2014;**15**:1484–90. https://doi.org/10.1021/bm500091e.
43. Maurizi MR. Proteases and protein degradation in *Escherichia coli*. *Experientia* 1992;**48**:178–201. https://doi.org/10.1007/BF01923511.
44. Zhang Y-Z, Ran L-Y, Li C-Y. *et al.* Diversity, structures, and collagen-degrading mechanisms of bacterial collagenolytic proteases. *Appl Environ Microbiol* 2015;**81**:6098–107. https://doi.org/10.1128/AEM.00883-15.
45. Xiang Z-X, Gong J-S, Shi J-H. *et al.* High-efficiency secretory expression and characterization of the recombinant type III human-like collagen in *Pichia pastoris*. *Bioresour Bioprocess* 2022;**9**:117. https://doi.org/10.1186/s40643-022-00605-4.
46. Raschmanová H, Weninger A, Knejzlík Z. *et al.* Engineering of the unfolded protein response pathway in *Pichia pastoris*: enhancing production of secreted recombinant proteins. *Appl Microbiol Biotechnol* 2021;**105**:4397–414. https://doi.org/10.1007/s00253-021-11336-5.
47. Mizuno K, Hayashi T, Bächinger HP. Hydroxylation-induced stabilization of the collagen triple helix: further characterization of peptides with 4-hydroxyproline in the xaa position. *J Biol Chem* 2003;**278**:32373–9. https://doi.org/10.1074/jbc.M304741200.
48. Schnicker NJ, Razzaghi M, Guha Thakurta S. *et al.* Bacillus anthracis prolyl 4-hydroxylase interacts with and modifies elongation factor Tu. *Biochemistry* 2017;**56**:5771–85. https://doi.org/10.1021/acs.biochem.7b00601.
49. Revell CK, Jensen OE, Shearer T. *et al.* Collagen fibril assembly: new approaches to unanswered questions. *Matrix Biol Plus* 2021;**12**:100079–94. https://doi.org/10.1016/j.mbplus.2021.100079.
50. Luo Y, Tao F, Wang J. *et al.* Development and evaluation of tilapia skin-derived gelatin, collagen, and acellular dermal matrix for potential use as hemostatic sponges. *Int J Biol Macromol* 2023;**253**:127014–25. https://doi.org/10.1016/j.ijbiomac.2023.127014.

51. Mathew-Steiner SS, Roy S, Sen CK. Collagen in wound healing. *Bioengineering* 2021;**8**:63–78. https://doi.org/10.3390/bioengineering8050063.
52. Ferreira AM, Gentile P, Chiono V. *et al.* Collagen for bone tissue regeneration. *Acta Biomater* 2012;**8**:3191–200. https://doi.org/10.1016/j.actbio.2012.06.014.
53. Sbricoli L, Guazzo R, Annunziata M. *et al.* Selection of collagen membranes for bone regeneration: a literature review. *Materials (Basel)* 2020;**13**:786–802. https://doi.org/10.3390/ma13030786.
54. Anishchenko I, Pellock SJ, Chidyausiku TM. *et al.* De novo protein design by deep network hallucination. *Nature* 2021;**600**:547–52. https://doi.org/10.1038/s41586-021-04184-w.
55. Dauparas J, Anishchenko I, Bennett N. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* 2022;**378**:49–56. https://doi.org/10.1126/science.add2187.
56. Repecka D, Jauniskis V, Karpus L. *et al.* Expanding functional protein sequence spaces using generative adversarial networks. *Nat Mach Intell* 2021;**3**:324–33. https://doi.org/10.1038/s42256-021-00310-5.
57. Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun* 2022;**13**:4348. https://doi.org/10.1038/s41467-022-32007-7.
58. O'Leary LER, Fallas JA, Bakota EL. *et al.* Multi-hierarchical self-assembly of a collagen mimetic peptide from triple helix to nanofibre and hydrogel. *Nat Chem* 2011;**3**:821–8. https://doi.org/10.1038/nchem.1123.
59. Jiang T, Vail OA, Jiang Z. *et al.* Rational Design of Multilayer Collagen Nanosheets with compositional and structural control. *J Am Chem Soc* 2015;**137**:7793–802. https://doi.org/10.1021/jacs.5b03326.
60. Wu J, Zhou L, Peng H. *et al.* A general and convenient peptide self-assembling mechanism for developing supramolecular versatile nanomaterials based on the biosynthetic hybrid amyloid-resilin protein. *Adv Mater* 2024;**36**:2304364–85. https://doi.org/10.1038/s41585-024-00962-z.
61. Safari B, Davaran S, Aghanejad A. Osteogenic potential of the growth factors and bioactive molecules in bone regeneration. *Int J Biol Macromol* 2021;**175**:544–57. https://doi.org/10.1016/j.ijbiomac.2021.02.052.
62. Li Y, Liu Y, Li R. *et al.* Collagen-based biomaterials for bone tissue engineering. *Mater Des* 2021;**210**:110049–72. https://doi.org/10.1016/j.matdes.2021.110049.
63. Wang X, Yu S, Sun R. *et al.* Identification of a human type XVII collagen fragment with high capacity for maintaining skin health. *Synth Syst Biotechnol* 2024;**9**:733–41. https://doi.org/10.1016/j.synbio.2024.06.001.
64. Žmitek K, Žmitek J, Rogl Butina M. *et al.* Effects of a combination of water-soluble coenzyme Q10 and collagen on skin parameters and condition: results of a randomised, placebo-controlled, double-blind study. *Nutrients* 2020;**12**:618–31. https://doi.org/10.3390/nu12030618.
65. Wu S, Letchworth GJ. High efficiency transformation by electroporation of *Pichia pastoris* pretreated with lithium acetate and dithiothreitol. *Biotechniques* 2004;**36**:152–4. https://doi.org/10.2144/04361DD02.
66. Huang P-S, Ban Y-EA, Richter F. *et al.* RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS One* 2011;**6**:e24109–17. https://doi.org/10.1371/journal.pone.0024109.
67. Abraham MJ, Murtola T, Schulz R. *et al.* GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 2015;**1**:19–25.
68. Gajula MNVP, Kumar A, Ijaq J. Protocol for molecular dynamics simulations of proteins. *Bio-protocol* 2016;**6**:e2051–62. https://doi.org/10.21769/BioProtoc.2051.
69. Tang M, Li T, Pickering E. *et al.* Steered molecular dynamics characterization of the elastic modulus and deformation mechanisms of single natural tropocollagen molecules. *J Mech Behav Biomed Mater* 2018;**86**:359–67. https://doi.org/10.1016/j.jmbbm.2018.07.009.
70. Tang M, Li T, Gandhi NS. *et al.* Heterogeneous nanomechanical properties of type I collagen in longitudinal direction. *Biomech Model Mechanobiol* 2017;**16**:1023–33. https://doi.org/10.1007/s10237-016-0870-6.
71. Maier JA, Martinez C, Kasavajhala K. *et al.* ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput* 2015;**11**(8):3696–713, DOI: https://doi.org/10.1021/acs.jctc.5b00255.
72. Hess B, Bekker H, Berendsen HJC. *et al.* LINCS: a linear constraint solver for molecular simulations. *J Comput Chem* 1997;**18**:1463–72. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.
73. Essmann U, Perera L, Berkowitz ML. *et al.* A smooth particle mesh Ewald method. *J Chem Phys* 1995;**103**:8577–93. https://doi.org/10.1063/1.470117.

Briefings in Bioinformatics, 2025, **26(1)**, bbae622
https://doi.org/10.1093/bib/bbae622
Problem Solving Protocol