Discovering design principles of collagen molecular stability using a genetic algorithm, deep learning, and experimental validation

PNAS RESEARCH ARTICLE | ENGINEERING
BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Discovering design principles of collagen molecular stability using a genetic algorithm, deep learning, and experimental validation

Eesha Khare[a,b,1], Chi-Hua Yu[a,c,1], Constancio Gonzalez Obeso[d,1], Mario Milazzo[a,e], David L. Kaplan[d], and Markus J. Buehler[a,f,g,2]

Edited by Yonggang Huang, Northwestern University, Glencoe, IL; received June 5, 2022; accepted August 18, 2022

Collagen is the most abundant structural protein in humans, providing crucial mechanical properties, including high strength and toughness, in tissues. Collagen-based biomaterials are, therefore, used for tissue repair and regeneration. Utilizing collagen effectively during materials processing ex vivo and subsequent function in vivo requires stability over wide temperature ranges to avoid denaturation and loss of structure, measured as melting temperature (T_m). Although significant research has been conducted on understanding how collagen primary amino acid sequences correspond to T_m values, a robust framework to facilitate the design of collagen sequences with specific T_m remains a challenge. Here, we develop a general model using a genetic algorithm within a deep learning framework to design collagen sequences with specific T_m values. We report 1,000 de novo collagen sequences, and we show that we can efficiently use this model to generate collagen sequences and verify their T_m values using both experimental and computational methods. We find that the model accurately predicts T_m values within a few degrees centigrade. Further, using this model, we conduct a high-throughput study to identify the most frequently occurring collagen triplets that can be directly incorporated into collagen. We further discovered that the number of hydrogen bonds within collagen calculated with molecular dynamics (MD) is directly correlated to the experimental measurement of triple-helical quality. Ultimately, we see this work as a critical step to helping researchers develop collagen sequences with specific T_m values for intended materials manufacturing methods and biomedical applications, realizing a mechanistic materials by design paradigm.

collagen | deep learning | thermal stability | generative algorithm | mechanics

Significance

Collagen is the most abundant structural protein in humans and as such, is often used in biomedical applications for tissue repair and regeneration. Designing de novo collagen to maintain its structural integrity in vivo, important for its mechanical performance and subsequent utility, remains a challenge today. In this work, we develop a deep learning framework to generate collagen sequences with desired thermal stability and validate our deep learning framework using both simulation and experiment. Given this validation, we discover key insights into the prevalence of amino acids in collagen triple helices and find a mechanistic relationship between our simulations and experiment. This framework enables researchers to develop collagen sequences with desired thermal stability for biomedical applications.

Collagen is the most abundant protein in animals and is found in the extracellular matrix of skin, tendons, bone, and vasculature as well as other tissues. The term "collagen" encompasses a family of at least 29 glycoproteins with common features responsible for the outstanding properties. Repeat units of glycine-X-Y (GXY) dominate the sequences, where the X and Y amino acids are usually occupied by proline (about 28% of the time) and hydroxyproline (about 38%) (1). These GXY sequences adopt a left-handed polyproline type II helical conformation, which after forming trimers, folds into a triple α-helical structure called tropocollagen, the basic structural unit of collagen (2). Tropocollagen is typically 300 nm long and 1.5 nm in diameter and assembles into hierarchical collagen structures including fibrils and fibers (Fig. 1*A*) (3–9). This hierarchical structure enables collagen to provide significant mechanical capacity under physiological conditions, exhibiting a tensile modulus of 0.2 to 0.86 GPa while maintaining elasticity in the human body (10–18).

Given this remarkable self-assembly process and the resulting mechanical properties, along with inherent biocompatibility, collagen-based biomaterials are routinely sought for in vivo tissue repairs, drug delivery systems, and other biomedical applications (19, 20). However, designing collagen to assemble in vitro to emulate the structural hierarchy and thermal stability of collagen in vivo remains challenging and limits the widespread use of collagen as biomaterial constructs. Therefore, most of the collagen used today has a reduced triple-helix content and thus, reduced thermal stability and mechanical properties, which result in rapid degradation in vivo. To overcome these challenges, synthetic collagen-based biomaterials are often stabilized via chemical crosslinking and related methods (21), which while effective in extending longevity in vivo, can negatively impact biological responses to collagen and alter the mechanical properties of the materials.

Given the importance of collagen's structural integrity for its mechanical function and thermal stability, one useful metric is the melting point (T_m), defined as the midpoint during the temperature window in which the collagen triple helix unfolds

Author contributions: E.K., C.-H.Y., C.G.O., D.L.K., and M.J.B. designed research; E.K., C.-H.Y., C.G.O., D.L.K., and M.J.B. performed research; E.K., C.-H.Y., C.G.O., M.M., D.L.K., and M.J.B. contributed new reagents/analytic tools; E.K., C.-H.Y., C.G.O., M.M., D.L.K., and M.J.B. analyzed data; and E.K., D.L.K., and M.J.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]E.K., C.-H.Y., and C.G.O. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: mbuehler@MIT.EDU.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2209524119/-/DCSupplemental.

Published September 26, 2022.

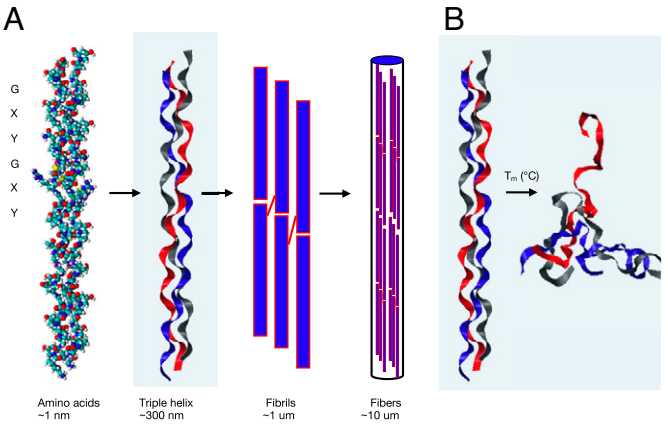

Fig. 1. The hierarchy of collagen helps maintain its structural integrity. (*A*) The collagen amino acid primary sequence, often in the form of G-X-Y repeat triplets, forms a larger chain. The three chains come together to form a triple helix, characteristic of collagen, which is also known as tropocollagen. The tropocollagen assembles into larger fibril and fiber units. (*B*) This work focuses on the thermal stability of tropocollagen. Thermal stability is characterized by the T_m value, which is the midpoint temperature of the denaturation process of the triple helix of tropocollagen to a disordered state. Once collagen is not in a triple helix, it no longer contributes to the mechanical stability of the larger fiber.

(Fig. 1*B*) (22). The thermal stability of collagen from different biological species or de novo collagen designs has been characterized experimentally (23–28). Computational studies have used molecular dynamics (MD) or coarse graining to determine sequence–structure–function thermal and mechanical properties in collagen across its different length scales (7, 12, 29, 30). These prior research efforts have made significant progress in understanding how mutations in the primary sequence affect the thermal stability of collagen. However, these approaches are computationally expensive and limited in the possibility to explore vast variations of sequences and mutations.

A predictive framework that facilitates the a priori design of collagen sequences with specific T_m values without prior knowledge of chemical interactions would enable the efficient design and subsequent synthesis of thermally stable collagens for specific applications. Such a framework for discovery could significantly propel the field of collagen-based biomaterials forward. Toward this goal, equations were developed to predict T_m values of collagen triple helices based on local interactions between different amino acid chemistries in collagen tripeptides following a GXY triplet ordering (31–33). In addition, an algorithm (scoring function for collagen-emulating peptides' temperature of transition) was developed to predict the registry and T_m values of synthetic collagen-based triple helices (34). Other recent approaches have been based on machine learning, which has emerged as a useful tool in the analysis of large datasets to help develop design principles for biological materials without knowledge of underlying biological interactions (35, 36). We recently published a deep learning–based framework trained on a large dataset of collagen sequences to quickly predict T_m values for a large number of mutations in collagen sequences (37).

Here, we report the development of a machine learning model to design de novo collagen sequences with desired T_m values. This approach builds on our previous deep learning algorithm, which applied a self-evolutionary algorithm, one-dimensional convolution, bidirectional long short-term memory (LSTM), and dropout features to predict T_m values of existing collagen sequences (37). To demonstrate the predictive power of our approach, we use MD, circular dichroism (CD) spectroscopy, and differential scanning calorimetry (DSC) to verify the T_m values of a few of our de novo collagen sequences. From this approach, we are able to derive two insights. First, our model has the highest predictive accuracy for de novo collagen sequences with strong triple-helix folding as measured through the triple-helical quality (ratio of positive to negative peak intensity [RPN]) value extracted from CD spectroscopy, and we demonstrate a correlation between hydrogen bonding in the triple helix found through MD and the RPN value (38). Second, given the high-throughput nature of our work, we identify key collagen triplet amino acid sequences that especially contribute to the thermal stability of collagen. These GXY triplet sequences should inform the design of the next generation of thermally stable collagen sequences. The goal of this work is to demonstrate the use of this generative algorithm in suggesting de novo collagen sequences with desired T_m values, thus contributing to a more efficient method of designing collagen-based materials with tailored properties.

Results

We report a generative model, implemented as a genetic algorithm, to generate collagen sequences (Fig. 2 and *SI Appendix*, Fig. S1) (39, 40). This model is named ColGen-GA to represent a collagen sequence generator, which is capable of generating (GA) homotrimeric type I collagen sequences with specific T_m values. ColGen-GA builds on our previous T_m predictor model ColGen (37), which uses a natural language processing method (41, 42). Amino acids within the collagen sequence were tokenized before passing them into the machine learning model (43, 44). Each collagen sequence is encoded with tokens, where amino acids are treated as a unique number from 1 to 21 so that the neural net can process the sequences.

Once collagen sequences are tokenized, they are passed into a genetic algorithm to generate new sequences inspired by the biological process of evolution, mimicking mutation, cross-over, and mating of chromosomes. Here, the collagen sequences serve as a chromosome, and the genes are represented as individual amino acids. Each generated collagen sequence is optimized to meet the objective function of the algorithm, which is a T_m value of choice. In this work, the T_m values are selected as 22 °C or room temperature and 37 °C or body temperature, as these are the two most relevant temperatures for bioengineering applications.

In the genetic algorithm, an initial population is randomly selected from the existing collagen dataset. Three parents are further randomly selected from the initial population to undergo tournament mating, where the two parents with the closest T_m values to the desired T_m value are selected. The T_m value is calculated from the previously reported ColGen model (37). These parents then undergo cross-over and mutation to produce "children" sequences. The cross-over and mutation rate are optimized to ensure that there is sufficient sampling of solutions, which prevents genetic drift while not leading to a loss of good solutions. This optimization is a balance between the number of generations required to reach convergence vs. the number of unique sequences generated (*SI Appendix*, Fig. S2). The child with the closest fit to the desired T_m value is then selected as the final output. This whole process is repeated over several iterations or "generations" until we reach a converged state around the desired T_m value (*SI Appendix*, Fig. S3). Further, we tested generation methods with "elitism," which is where the best children are overrepresented in the initial population such that the better traits stay in the genetic pool for longer, and "randomness," which is where the initial

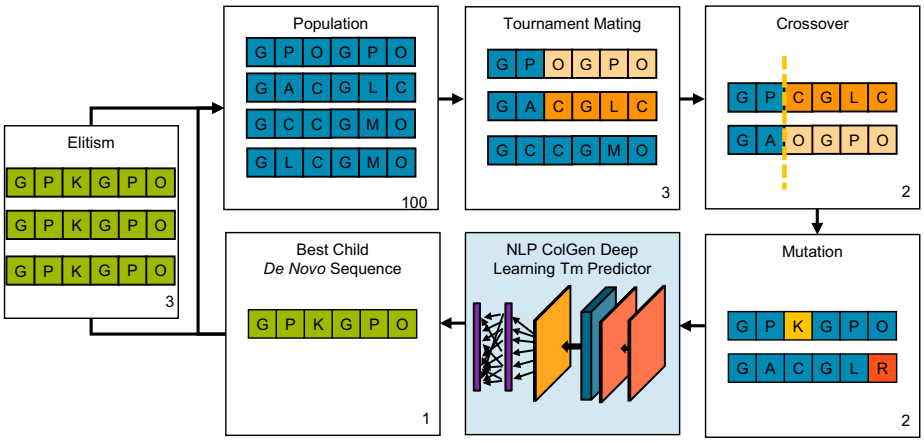

Fig. 2. The machine learning–based genetic algorithm used. Three sequences are randomly selected from a randomly generated population based on the dataset collagen sequences. The three sequences undergo tournament mating to identify the two best parents; these parents undergo further cross-over and mutations within their sequences to produce children offspring. The resulting children are then evaluated with the NLP (natural language processing) ColGen deep learning T_m predictor, and the best child matching the desired T_m value objective function is output. If elitism is implemented in the model, the child is overrepresented in the initial population to help preserve its general sequence features. Numbers in the bottom right of the boxes represent the numbers of sequences in each stage.

population in the next generation is unrelated to the children from the previous generation. While elitism helps ensure that the quality of the generative algorithm does not decrease over time, it has a disadvantage of converging on a local minimum rather than finding the best solution. Selecting appropriate population sizes, mutations and cross-over frequencies, and elitism is critical in the algorithm's ability to generate a wide number of sequences without losing the best features found in the generations. The specific parameters used in this model are listed in *SI Appendix*, Table S1. Due to computational modesty, the genetic algorithm model can easily be deployed on a laptop or desktop computer without further requirement of GPUs.

The ColGen-GA model can quickly reach the desired T_m value and maintain that value over several generations. In *SI Appendix*, Fig. S3, the desired normalized T_m value of 0.9, corresponding to 37 °C, is reached almost immediately. The convergence is even faster when elitism is implemented.

From the ColGen-GA, we are able to produce several de novo sequences within our desired temperature range (T_m values of 22 °C and 37 °C). We selected two sequences in each temperature for further validation (collagen-like peptide 1 [CP1] and CP3 for 22 °C, CP2 and CP4 for 37 °C) as well as another de novo sequence generated from a different generative algorithm where the collagen primary sequence is not required to be in a G-X-Y order (CP5). Table 1 shows strong agreement between our initial prediction of the T_m value and the T_m value found using an experiment within a few degrees centigrade. Temperature sweep experiments revealed that the T_m values for the CPs and type I collagen control as measured by CD and DSC were in good agreement with those predicted by ColGen-GA (Fig. 3, Table 1, and *SI Appendix*, Fig. S4). The slight difference between CD and DSC is attributed to the higher heating rate in DSC experiments.

The CD spectra of the de novo CPs show that the CPs are able to form triple-helical structures (Fig. 4 *A* and *B*). The CPs and the control follow a standard CD triple helix–forming collagen spectrum. There is a clear positive signal at 222 nm in the 5 °C wavelength scan (Fig. 4 *A* and *B*), related to the

Table 1. Summary of the names, amino acid sequences, and T_m values of samples studied listed in order of increasing to decreasing T_m value

Name	Sequence	Method	ColGen T_m (°C) model	CD T_m (°C) EXP	DSC T_m (°C) EXP	T_m calculator* (°C) model
Collagen type 1	Bovine collagen	Control	—	40.6	40.9	—
Std.	GPOGPOGPOGPOGPOGPOGPOGPOGPO	Reference	62.0	—	—	63.8
CP5	GPOGPOGPOGPOGPOGPPAGPOGROGRO	Previous algorithm†	46.6	21.5, 40.4, 60.9	26.5, 44.9, 62.8	22.2‡
CP4	GYOGPOGPOGKOGPOGKOGPOGPOGPHGPM	Random	37.7	41.2	42.8	40.6
CP2	GPOGPOGPRGMOGPOGPOGPOGPO	Elitism	37.3	35.4	36.4	38.5
CP3	GPOGPOGDOGATGPOGRCGPQGPOGPOGPO	Elitism	22.0	20.8	22.6	21.1
CP1	GIAGPAGPOGDAGPOGPOGPOGPO	Random	22.2	18.6	20.4	25.0
CP6	GVMGWGGALGYHGERGMNGHTGND	Previous algorithm†	−3.3	Does not form a stable helix		−76.2
CP7	GEIGEVGSHGVNGHEGGFGYGGMGGG	Previous algorithm†	−26.6	Does not form a stable helix		−83.0

EXP, experimental measurement.
*The T_m calculator prediction is from the work of Persikov et al. (22, 31–33, 71, 72).
†These de novo collagen peptides were generated from a previous genetic algorithm not discussed in the paper. Their T_m values, however, were predicted from ColGen, and as such, they are useful for understanding the validation of the ColGen model.
‡The T_m calculator prediction from the work of Persikov et al. (22, 31–33, 71, 72) is unable to calculate T_m values for sequences that do not follow $(GXY)_n$ formatting.

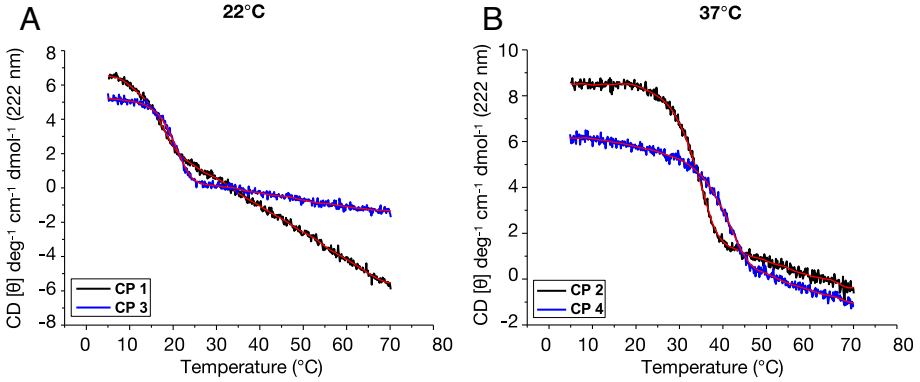

Fig. 3. CD temperature scan at 222 nm for collagen peptides demonstrating triple-helix structure: (*A*) 22 °C peptides CP1 and CP3 and (*B*) 37 °C peptides CP1 and CP3. Scans at 1 °C/min with sampling every 0.1 °C indicate that de novo peptides have T_m values within a couple of degrees of the target T_m.

presence of a triple helix, which disappears in the 70 °C scan in which collagen denatures (Fig. 4 *A* and *C*). The ratio of positive signal at 222 nm to negative at 196 to 198 nm (RPN) serves as a concentration-independent measurement of the quality and quantity of triple-helix formation. An RPN value of 0.133 for type I collagen is the highest compared with all the CPs (*SI Appendix*, Table S2), indicating that it forms the best triple helix; as expected, CP4 exhibited the highest RPN (0.129) of all four CPs, followed by CP2 (0.121) and CP3 (0.104), values similar to those of type I collagen. In contrast, CP1 and CP5 exhibited RPN values 61.6 and 35.3% lower than the control, reaching values of 0.086 and 0.055, respectively. These results indicate that CP2 to -4 were able to interact cooperatively, developing stable triple helixes in a similar way to type I collagen. Such interactions were less favorable for CP1 and CP5, as seen by the lower RPN. The RPN value follows the order of type I collagen control ~ CP4 > CP2 > CP3 > CP5 > CP1. This is also consistent with the intended T_m values of the CPs, where CP2 and CP4 were designed to have higher T_m values (around 37 °C) and thus, maintain a more stable triple-helical configuration. Interestingly, CP5 showed a multistep denaturation process with temperature, which was related to the interrupted GXY sequence (*SI Appendix*, Fig. S5). This multistep behavior hinders the assignment of a single T_m value to CP5.

Upon experimentally measuring the RPN and the T_m, we found that the higher the RPN value, the lower the differences between ColGen-GA predicted and measured T_m values (Fig. 5*A*). This is likely because a higher RPN corresponds to a higher-quality triple helix, which is more likely present for the high–T_m value sequences as discussed.

To further validate the ColGen-GA model and provide support to the CD and DSC experiments, MD simulations were also used to simulate experimental heating of the triple-helical peptides. While exact T_m cannot be extracted from MD due to the faster heating rate used in simulation compared with experiment, MD simulations confirm that the stability ordering of the CPs is $(GPO)_{10}$ > CP4 ~ CP2 > CP3 > CP5 > CP1 (Fig. 5*B*), as observed in CD and DSC measurements. In MD, $(GPO)_{10}$ rather than bovine collagen is used as a model collagen peptide mimetic with the highest T_m value. The MD results demonstrate that all of the peptides are correctly ordered in terms of their thermal stabilities, except for CP5, whose thermal stability in MD simulations is predicted to be much less than experimentally measured. This discrepancy is likely due to a poorer prediction of triple-helical structure for CP5, as it does not follow the GXY pattern consistently.

MD simulations also enabled us to further validate the relationship between RPN and T_m values and the accuracy of our predictions by developing a mechanistic understanding of the different CPs. We evaluated the triple-helix quality of the different CPs by measuring the amount of hydrogen bonds between the strands as a proxy for triple-helix strength and related it with their RPN value (*SI Appendix*, Fig. S6). We found that the CPs with higher RPN present more hydrogen bonds in their triple-helix structure compared with CPs with lower RPN values (Fig. 5*C* and *SI Appendix*, Table S3). An

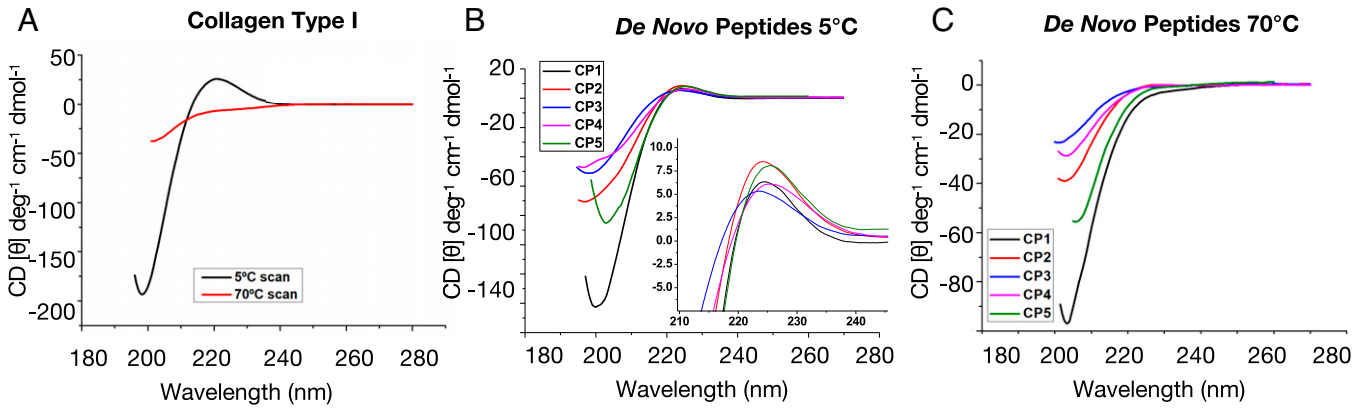

Fig. 4. CD wavelength scan at 222 nm for collagen peptides demonstrating triple-helix structure: (*A*) type I collagen as a control at 5 °C and 70 °C and de novo peptides at (*B*) 5 °C and (*C*) 70 °C. (*B*) *Inset* is zoomed into wavelength ranges from 210 to 250 nm for clarity. De novo peptides demonstrate the same characteristic behavior as the type I collagen sequence, indicating that they have a triple-helical structure. Both the type I collagen and de novo peptides denature at 0 °C.

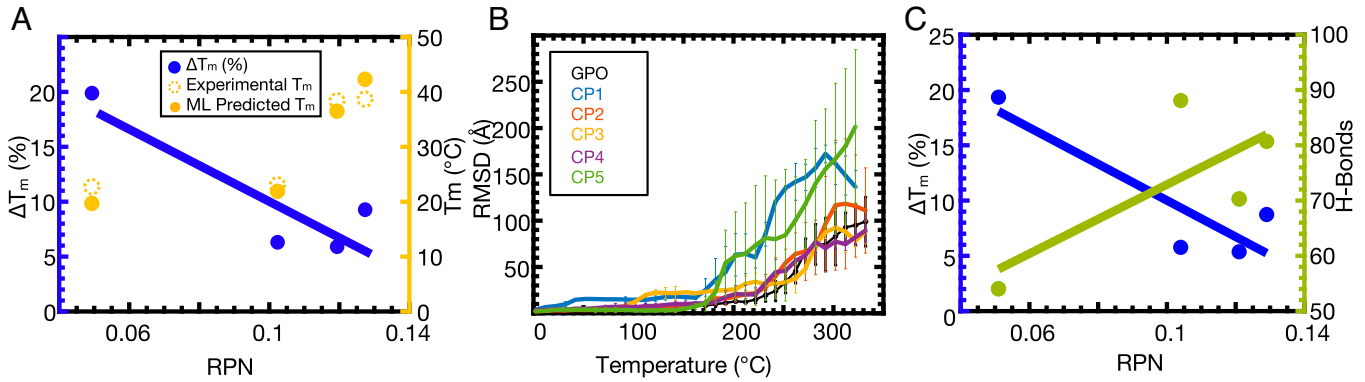

Fig. 5. Relationship between collagen triple-helix quality and T_m values using experiment and MD simulation. (*A*) There is an inverse relationship between the RPN and the difference in T_m value between the experimental CD T_m and ColGen machine learning (ML) predicted T_m: $\Delta T_m = \frac{(T_{m,\ experiment\ CD} - T_{m,\ ColGen\ -\ GA})}{T_{m,\ experiment\ CD}}$. This indicates that the ColGen algorithm is able to more robustly predict the thermal stability of higher-quality triple helices. The RPN also follows a direct relationship with T_m value, indicating that more stable triple helices have a higher T_m. (*B*) MD simulations show that the CPs maintain roughly the expected stability, measured by rmsd of the triple helix as predicted by ColGen. (*C*) Hydrogen bonding analysis at 50 °C in the MD simulation shows a similar correlation as the RPN in *A*. Peptides with more hydrogen bonding generally have a lower deviation from ColGen-predicted T_m values compared with experimental T_m values. Further, RPN has a direct relationship with the number of hydrogen bonds in the CP, indicating that a higher-quality triple helix has more hydrogen bonding.

image of representative hydrogen bonds is provided in *SI Appendix*, Fig. S7, and we note that in our sequences, glycine qualitatively demonstrates the most hydrogen bonding. To our best knowledge, this is a demonstration of the direct relationship between the number of hydrogen bonds computed in MD and RPN values experimentally measured with CD.

Given the validation of the model with experiment and MD, we conducted high-throughput processing to derive insights into GXY triplets of collagen that are most suitable in achieving desired T_m values. After generating 1,000 de novo collagen sequences with T_m values of 22 °C and 37 °C (Datasets S1 and S2 and *SI Appendix*, Fig. S8), we found the top 1.3% most commonly occurring triplets within our generated sequences and determined their co-occurrence matrices in Fig. 6 *A* and *B*. The co-occurrence matrix helps show which GXY triplets occur with other GXY triplets to provide a graphical insight to how to build a larger sequence from a combination of triplets. GPO emerges as the most commonly present triplet in the generated sequences. This is in agreement with the literature because GPO is the canonical triplet in increasing the strength of CPs, and $(GPO)_x$ is often used as a gold standard in collagen mimetic peptides for thermal stability (33). Beyond the presence of GPO, we also find a number of other triplets that emerge as useful motifs in achieving the desired T_m values. In alignment with others who have noted the stabilizing effect of KGE/KGD (33, 34, 45), lysine, glutamic acid, and aspartic acid contribute stability to the collagen peptide, as the residues GPK, GEO, and GDO have a minimal decrease in thermal stability compared with other frequently occurring triplets (Fig. 6*B*). Interestingly, all of these triplets are in a GPY or GXO

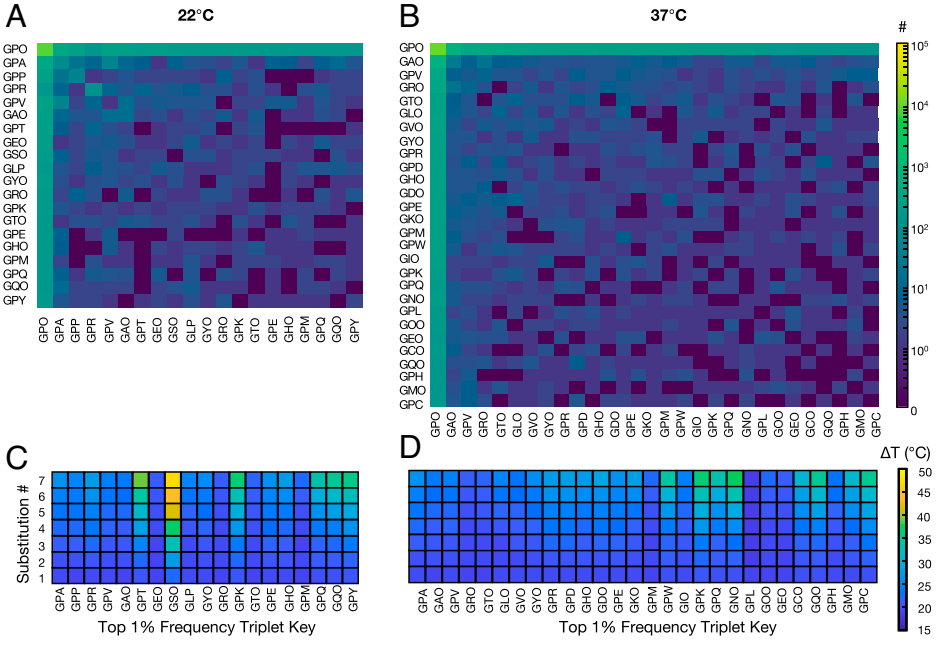

Fig. 6. High-throughput identification of the most frequent sequences in de novo collagen peptides. The co-occurrence matrix of the 1,000 generated de novo collagen sequences for 22 °C (*A*) and 37 °C (*B*) when sorted by the most frequent triplets shows which triplets occur together in the same sequence. These most frequent triplets from 22 °C (*C*) and 37 °C (*D*) are substituted *n* times into a $(GPO)_{14}$ ideal standard peptide, and their destabilizing effect on T_m is evaluated, where $\Delta T_m = T_m(GPO)_{14} - T_m(sequence)$.

configuration, where either P or O is present with the other guest amino acid replaced. Further, these guest amino acids do not follow a consistent physicochemical trend, and there is a range of hydrophobic, polar, or charged residues that contribute to the mechanical stability of the CP. This lack of physicochemical consistency is not something we could have inferred from analytical outcomes given that there is no trend in the data. We also note that the demonstrated guest amino acids do not follow the same occurrence as amino acids naturally occurring in collagen, where basic residues are in the Y position and Glu or hydrophobic residues are in the X position (33). Instead, the guest amino acids demonstrated here follow a different occurrence given that this is a synthetic set. This provides further justification for using large data generation means through machine learning to derive design principles. Further, comparing the highest-frequency triplets in the 22 °C (Fig. 6*A*) to 37 °C (Fig. 6*B*) data, we find that ∼50% of the most frequent triplets in 22 °C are also found in the list of most frequent triplets at 37 °C. These are likely the triplets that contribute most to the thermal stability of the collagen sequence, while the other triplets help achieve the lower target temperature of 22 °C.

Given these most frequent triplets, we sought to understand which triplets had the greatest effect on the T_m values (Fig. 6 *C* and *D*). When substituted into the highest-stability $(GPO)_{14}$ sequence, the top triplets in the 37 °C sequences induce a lower amount of destabilization of the T_m value, where the destabilization is measured as $\Delta T_m = T_m(GPO)_{14} - T_m$ (de novo sequence), compared with the most frequent triplets in the 22 °C de novo sequences.

Discussion

We developed a platform that uses a deep learning model trained with input sequences from the literature to generate de novo collagen sequences with desired T_m values. The model, ColGen-GA, incorporates an LSTM-based T_m predictor reported in our previous work (37) and a generative genetic algorithm to produce new sequences with specified thermal stability behavior. We then validated our model by selecting some of the generated sequences and testing them experimentally with CD and DSC and computationally with MD simulations. The CD experiments confirmed that these new collagen mimetic peptides had triple-helical structure, and together with DSC, the experiments confirmed the predicted T_m values of the de novo sequences within a few degrees centigrade. By studying the quality of the triple-helical formation of the CPs, as measured through its RPN value, we determined that higher–RPN value CPs have less deviation between the experimental and ColGen-GA–predicted T_m values. This means that CPs that have a higher-quality triple helix and thus, a higher T_m are likely to have a better T_m prediction than CPs that are less able to form triple helices. These experimental results were also further validated by MD simulations, which showed that the de novo peptides followed the stability order as predicted and that more hydrogen bonds correlate directly with higher RPN values and higher T_m values. Such a result is particularly relevant since a mechanistic understanding of the relationship between the RPN values and the number of H bonds promotes a deeper understanding and rationalization of the thermal stability of collagen and more broadly, protein sequences.

Given the validation of the model, we used the large dataset and computational power of the machine learning model to discover important triplets in the collagen sequences. ColGen-GA enables the fast generation and prediction of T_m values of 1,000 new sequences in just 8 h using a laptop, compared with the 10 d per collagen sequence simulation required for MD on two nodes and 32 CPUs. Because of this computational power and speed, we are able to derive insights into important collagen sequences compared with what our previous modeling or experimental capacities enabled. We identified the highest-frequency GXY triplets from the de novo sequences generated for target T_m values of 22 °C and 37 °C. The triplets identified can be used by other researchers when designing new collagen sequences with specific T_m values. To assess these triplets in the context of a longer sequence, we also provided a co-occurrence map to understand how these important triplets work together in longer sequences. These triplets would not have been discovered through analytical means, as we find that a consistent physicochemical principle, such as hydrophobicity or charge, to explain the triplet behavior is not present in the most frequent triplets. We envision that the triplets identified here could be used in creating useful collagen sequences.

While the model enables the rapid generation of sequences at desired T_m values, there are some limitations that can be addressed in future work. These limitations can primarily be addressed by expanding the dataset to include more collagen sequences with varying lengths and compositions. Most of the collagen sequences used in our present dataset were collagen mimetic peptides rather than complete collagen sequences. Incorporating longer sequences would help in the design of collagen proteins from bacteria, which often produce longer protein sequences (26, 46, 47). Further, most of the sequences in the dataset incorporate hydroxyproline (O) or a specific subset of guest residues. As such, building sequences that do not incorporate O, which is an especially important limitation to bacterially produced collagens that have no means of producing O, would be challenging with the current dataset used. Expanding the compositional diversity of the dataset would also help further improve the prediction. While we have strong predictive capacity (*SI Appendix*, Figs. S9 and S10), this predictive capacity decreases at the lower and higher ranges of Tm values, where fewer data points exist in the training dataset. Adding more triplet sequence variety would improve this constantly evolving machine learning model. Another limitation of exploiting GA could be the efficiency. Compared with other optimization tools, such as the gradient method, GA has a slow computational speed when processing large amounts of initial populations or local minima. The convergence rate for the same size of the initial population can be modulated by simulated annealing to change the probability of cross-over and mutation on the fly (48). Finally, manipulating the model itself would help with the prediction of heterotrimer sequences beyond the homotrimers presented here.

Despite these limitations, the reported approach represents a powerful and efficient tool in the design of collagen sequences with specific T_m values. Our approach should lead to the design of collagen biomaterials and tunable properties with a priori–desired T_m values. Further, our presentation of triplets will help inform how to build mechanically robust collagen sequences at desired temperatures into the future, especially given the vast design space of 10^{21} combinations of $(GXY)_{10}$ sequences.

Beyond the generation of sequences as an engineering tool, our approach contributes to an understanding of collagen denaturation rates and how these T_m values correlate to structure. Such information is important (for example, in understanding the mechanical behavior of specific tissues with impacts on

denaturation or biological function in scenarios, such as thermal treatments for cancer). Further, many collagen-based diseases, such as *Osteogenesis imperfecta*, are based on mutations in the primary sequence of collagen. This method would help offer insight and perspectives on these disease states in the context of thermal stability, with implications for future repair routes. Another aspect is collagen degradation by matrix metalloproteinases (MMPs), crucial in many physiological processes, such as wound healing, tissue remodeling, and organ morphogenesis. It is well known that stable triple helices are far more resistant to MMP degradation than denatured collagen, reflective of the structural stability of the matrices (49, 50). Thus, CPs capable of forming better triple helices are, thus, more resistant to MMP degradation. Similarly, higher mechanical integrity and structural order of collagen result in a more robust collagen matrix (49, 51). Considering that human mesenchymal stem cells proliferate, propagate, and differentiate in response to the mechanical properties of the matrix they develop in refs. 52–54, we envision that the ability of designing collagen sequences with tailored thermal stability with this deep learning method would allow us to create biomaterials with on-demand MMP degradation rates, mechanical properties, and tailored influence on cell behavior. Finally, the role of collagen sequences in the context of mineralization in vivo, such as with hydroxyapatite and bone formation, can benefit from these methods related to engineering approaches to modulate organic (collagen) and inorganic (e.g., hydroxyapatite) interfaces related to mechanics and bone structure–function.

Materials and Methods

Collagen Dataset. We collected 633 homotrimer collagen sequences with reported T_m values from a survey of the literature (Dataset S3) (4, 31, 33, 47, 49, 55–72). The distribution of the dataset is presented in *SI Appendix*, Fig. S6, where sequences have experimentally measured melting temperatures ranging from −17 °C to 70 °C, with a mean value at ∼30 °C. The data show a normal distribution (*SI Appendix*, Fig. S10). The dataset is used to train the deep learning ColGen model.

Collagen Samples. Several de novo sequences from the ColGen-GA model were selected for synthesis and experimental validation. These CPs were synthetized by GeneScript Biotech (95% purity and trifluoroacetic acid removal). As the triple helix–forming control, commercially available bovine type I collagen was used (PureCol TypeI Collagen, catalog no. 5005; AdvancedBiomatrix). Table 1 summarizes the peptide naming scheme, amino acid composition, and experimentally measured T_m values.

CD. Spectra were acquired using a Jasco J-815 Circular Dichroism Spectrometer. CPs and bovine collagen type I were dissolved in PBS at 0.3 mg/mL (final pH of 7.1 to 7.3). Samples were kept at 5 °C for 72 h before scanning in the far ultraviolet (UV) (180 to 260 nm) at 5 °C. Ellipticity at 222 nm was monitored as a function of temperature while heating the samples from 5 °C to 70 °C at 1 °C/min with data collection every 0.1 °C. For derivatization of the temperature scans and calculation of the minimum of the first derivative, the data were smoothed using a fast Fourier transform filter with a cutoff frequency of 0.342 Hz. T_m values were calculated as the minimum of the first derivative of the temperature scans. After reaching 70 °C, the temperature was maintained, and samples were scanned from 180 to 260 nm. CD spectra included accumulating three scans at a scanning speed of 20 nm/s and 4 s of digital integration time. For all plotted data, the high-tension voltage of the photomultiplier was kept below 600 V.

DSC. Thermograms were acquired using a TA Instruments DSC (Q100 series; TA Instruments). CPs and collagen type I were dissolved in phosphate buffered saline (PBS) at 50 mg/mL (final pH of 7.1 to 7.4) and kept at 5 °C for 72 h before measurement. A total of 20 μL of each sample was hermetically sealed in an aluminum pan (Hermetic Tzero pans model 901684.901; TA Instruments) and scanned from 5 °C to 65 °C at a rate of 2.5 °C/min using as a reference 20 μL of PBS. The melting temperature was considered as the minimum of the endotherm (*SI Appendix*) (32).

MD. MD simulations were performed using the Nanoscale Molecular Dynamics (NAMD) code with the Chemistry at Harvard Macromolecular Mechanics (CHARMM) force field (73, 74), which also includes parameters for the hydroxyproline residue. We prepared each peptide topology using the triple-helical collagen building script (75) based on the primary amino acid composition, including the hydroxyproline residue. The protein was solvated with a 2.4-nm boundary water box using TIP3P (transferable intermolecular potential with 3 points) water molecules as the solvent. The total number of atoms in the solvated system was ∼90,000. A 1-fs time step was used, and rigid bonds were applied to constrain the bonds of the water molecules. van der Waals interactions were computed using a cutoff for a neighbor list at 1.4 nm, with a switching function from 1.0 to 1.2 nm. For electrostatic interactions, the particle-mesh Ewald sums method was used with periodic boundary conditions. A preliminary energy minimization was performed using a steepest descent algorithm. The systems were then equilibrated at 275 K for 5 ns each under a constant atom, volume, and temperature (NVT) and then, constant pressure (NPT) ensemble. The resulting systems were further equilibrated under NVT for 2 ns before beginning the heating process to mimic the CD and DSC experiments. The temperature of the simulation was increased by 10 K every 10 ns from 275 to 600 K (45). rmsd of the protein backbone and hydrogen bonding number were determined from the last 6 ns of each temperature by using visual MD plug-ins on trajectory files that were output every 50 ps. Each simulation was repeated three times.

Data, Materials, and Software Availability. All study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. This research was supported by the MIT-IBM Watson AI Lab, Army Research Office (ARO) Grant W911NF-17-1-0384, and NIH Grants P41EB027062 and U01 EB014976. E.K. acknowledges support from NSF Graduate Research Fellowship Program (GRFP). C.-H.Y. acknowledges support from Ministry of Science and Technology in Taiwan Grant MOST 109-2222-E-006-005-MY2. M.J.B. acknowledges support from Office of Naval Research (ONR) Grants N000141612333 and N000141912375. We acknowledge fruitful discussions with Barbara Brodsky.

Author affiliations: [a]Laboratory for Atomistic and Molecular Mechanics, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [c]Department of Engineering Science, National Cheng Kung University, Tainan 701, Taiwan; [d]Department of Biomedical Engineering, Tufts University, Medford, MA 02155; [e]Department of Civil and Industrial Engineering, University of Pisa, 56122 Pisa, Italy; [f]Center for Computational Science and Engineering, Schwarzman College of Computing, Massachusetts Institute of Technology, Cambridge, MA 02139; and [g]Center for Materials Science and Engineering, Cambridge, MA 02139

1. A. Sorushanova *et al.*, The collagen suprafamily: From biosynthesis to advanced biomaterial development. *Adv. Mater.* **31**, e1801651 (2019).
2. H. Lodish, A. Berk, S. L. Zipursky, *Molecular Cell Biology: The Fibrous Proteins of the Matrix* (W. H. Freeman, 2000).
3. D. J. Prockop, K. I. Kivirikko, Collagens: Molecular biology, diseases, and potentials for therapy. *Annu. Rev. Biochem.* **64**, 403–434 (1995).
4. J. P. R. O. Orgel, T. C. Irving, A. Miller, T. J. Wess, Microfibrillar structure of type I collagen in situ. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 9001–9005 (2006).
5. G. N. Ramachandran, G. Kartha, Structure of collagen. *Nature* **176**, 593–595 (1955).
6. A. Rich, F. H. Crick, The structure of collagen. *Nature* **176**, 915–916 (1955).
7. M. J. Buehler, S. Y. Wong, Entropic elasticity controls nanomechanics of single tropocollagen molecules. *Biophys. J.* **93**, 37–43 (2007).
8. A. Bhattacharjee, M. Bansal, Collagen structure: The Madras triple helix and the current scenario. *IUBMB Life* **57**, 161–172 (2005).
9. R. Puxkandl *et al.*, Viscoelastic properties of collagen: Synchrotron radiation investigations and structural model. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **357**, 191–197 (2002).
10. R. K. Nalla, J. J. Kruzic, J. H. Kinney, R. O. Ritchie, Mechanistic aspects of fracture and R-curve behavior in human cortical bone. *Biomaterials* **26**, 217–231 (2005).
11. R. O. Ritchie, J. J. Kruzic, C. L. Muhlstein, R. K. Nalla, E. A. Stach, Characteristic dimensions and the micro-mechanisms of fracture and fatigue in "nano" and "bio" materials. *Int. J. Fract.* **128**, 1–15 (2004).
12. M. J. Buehler, Nature designs tough collagen: Explaining the nanostructure of collagen fibrils. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 12285–12290 (2006).
13. A. Gautieri, S. Vesentini, A. Redaelli, M. J. Buehler, Viscoelastic properties of model segments of collagen molecules. *Matrix Biol.* **31**, 141–149 (2012).

14. Z. L. Shen, M. R. Dodge, H. Kahn, R. Ballarini, S. J. Eppell, Stress-strain experiments on individual collagen fibrils. *Biophys. J.* **95**, 3956–3963 (2008).
15. J. A. J. van der Rijt, K. O. van der Werf, M. L. Bennink, P. J. Dijkstra, J. Feijen, Micromechanical testing of individual collagen fibrils. *Macromol. Biosci.* **6**, 697–702 (2006).
16. L. Yang *et al.*, Mechanical properties of native and cross-linked type I collagen fibrils. *Biophys. J.* **94**, 2204–2211 (2008).
17. R. B. Svensson, T. Hassenkam, C. A. Grant, S. P. Magnusson, Tensile properties of human collagen fibrils and fascicles are insensitive to environmental salts. *Biophys. J.* **99**, 4020–4027 (2010).
18. M. Milazzo, G. S. Jung, S. Danti, M. J. Buehler, Wave propagation and energy dissipation in collagen molecules. *ACS Biomater. Sci. Eng.* **6**, 1367–1374 (2020).
19. C. H. Lee, A. Singla, Y. Lee, Biomedical applications of collagen. *Int. J. Pharm.* **221**, 1–22 (2001).
20. R. Parenteau-Bareil, R. Gauvin, F. Berthod, Collagen-based biomaterials for tissue engineering applications. *Materials (Basel)* **3**, 1863–1887 (2010).
21. M. Milazzo *et al.*, Additive manufacturing approaches for hydroxyapatite-reinforced composites. *Adv. Funct. Mater.* **29**, 1903055 (2019).
22. A. V. Persikov, J. A. M. Ramshaw, A. Kirkpatrick, B. Brodsky, Electrostatic interactions involving lysine make major contributions to collagen triple-helix stability. *Biochemistry* **44**, 1414–1422 (2005).
23. T. V. Burjanadze, Hydroxyproline content and location in relation to collagen thermal stability. *Biopolymers* **18**, 931–938 (1979).
24. K. Gekko, S. Koga, Increased thermal stability of collagen in the presence of sugars and polyols. *J. Biochem.* **94**, 199–205 (1983).
25. B. J. Rigby, Amino-acid composition and thermal stability of the skin collagen of the Antarctic ice-fish. *Nature* **219**, 166–167 (1968).
26. A. Mohs *et al.*, Mechanism of stabilization of a bacterial collagen triple helix in the absence of hydroxyproline. *J. Biol. Chem.* **282**, 29757–29765 (2007).
27. K. Inouye *et al.*, Synthesis and physical properties of (hydroxyproline-proline-glycine)10: Hydroxyproline in the X-position decreases the melting temperature of the collagen triple helix. *Arch. Biochem. Biophys.* **219**, 198–203 (1982).
28. S. Sakakibara *et al.*, Synthesis of (Pro-Hyp-Gly)n of defined molecular weights. Evidence for the stabilization of collagen triple helix by hydroxypyroline. *Biochim. Biophys. Acta* **303**, 198–202 (1973).
29. G. Gronau *et al.*, A review of combined experimental and computational procedures for assessing biopolymer structure-process-property relationships. *Biomaterials* **33**, 8240–8255 (2012).
30. M. J. Buehler, Atomistic and continuum modeling of mechanical properties of collagen: Elasticity, fracture, and self-assembly. *J. Mater. Res.* **21**, 1947–1961 (2006).
31. A. V. Persikov, J. A. M. Ramshaw, A. Kirkpatrick, B. Brodsky, Amino acid propensities for the collagen triple-helix. *Biochemistry* **39**, 14960–14967 (2000).
32. A. V. Persikov, J. A. M. Ramshaw, A. Kirkpatrick, B. Brodsky, Peptide investigations of pairwise interactions in the collagen triple-helix. *J. Mol. Biol.* **316**, 385–394 (2002).
33. A. V. Persikov, J. A. M. Ramshaw, B. Brodsky, Prediction of collagen stability from amino acid sequence. *J. Biol. Chem.* **280**, 19343–19349 (2005).
34. D. R. Walker *et al.*, Predicting the stability of homotrimeric and heterotrimeric collagen helices. *Nat. Chem.* **13**, 260–269 (2021).
35. J. Wang, H. Cao, J. Z. H. Zhang, Y. Qi, Computational protein design with deep learning neural networks. *Sci. Rep.* **8**, 6349 (2018).
36. A. Al-Shahib, R. Breitling, D. R. Gilbert, Predicting protein function by machine learning on amino acid sequences–a critical evaluation. *BMC Genomics* **8**, 78 (2007).
37. C. H. Yu *et al.*, ColGen: An end-to-end deep learning model to predict thermal stability of de novo collagen sequences. *J. Mech. Behav. Biomed. Mater.* **125**, 104921 (2022).
38. E. S. Hwang, G. Thiagarajan, A. S. Parmar, B. Brodsky, Interruptions in the collagen repeating tripeptide pattern can promote supramolecular association. *Protein Sci.* **19**, 1053–1064 (2010).
39. D. Whitley, A genetic algorithm tutorial. *Stat. Comput.* **4**, 65–85 (1994).
40. S. Katoch, S. S. Chauhan, V. Kumar, A review on genetic algorithm: Past, present, and future. *Multimedia Tools Appl.* **80**, 8091–8126 (2021).
41. J. Pennington, R. Socher, C. D. Manning, "GloVe: Global vectors for word representation" in *Proceedings of the Conference on EMNLP 2014–2014 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2014), pp. 1532–1543.
42. C. Mirabello, B. Wallner, rawMSA: End-to-end deep learning using raw multiple sequence alignments. *PLoS One* **14**, e0220182 (2019).
43. L. Qin, G. Dong, J. Peng, "Chemical-protein interaction extraction via ChemicalBERT and attention guided graph convolutional networks in parallel" in *Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020* (Institute of Electrical and Electronics Engineers Inc., 2020), pp. 708–715.
44. X. Li, D. Fourches, SMILES pair encoding: A data-driven substructure tokenization algorithm for deep learning. *J. Chem. Inf. Model.* **61**, 1560–1569 (2021).
45. N. Keshwani, S. Banerjee, B. Brodsky, G. I. Makhatadze, The role of cross-chain ionic interactions for the stability of collagen model peptides. *Biophys. J.* **105**, 1681–1688 (2013).
46. H. Cheng *et al.*, Location of glycine mutations within a bacterial collagen protein affects degree of disruption of triple-helix folding and conformation. *J. Biol. Chem.* **286**, 2041–2046 (2011).
47. Z. Yu, B. An, J. A. M. Ramshaw, B. Brodsky, Bacterial collagen-like proteins that form triple-helical structures. *J. Struct. Biol.* **186**, 451–461 (2014).
48. K. N. Brown, J. Cagan, Optimized process planning by generative simulated annealing. *Artif. Intell. Eng. Des. Anal. Manuf.* **11**, 219–235 (1997).
49. M. D. Shoulders, R. T. Raines, Collagen structure and stability. *Annu. Rev. Biochem.* **78**, 929–958 (2009).
50. M. W. H. Kirkness, N. R. Forde, Single-molecule assay for proteolytic susceptibility: Force-induced collagen destabilization. *Biophys. J.* **114**, 570–576 (2018).
51. A. Terzi *et al.*, Effects of processing on structural, mechanical and biological properties of collagen-based substrates for regenerative medicine. *Sci. Rep.* **8**, 1429 (2018).
52. J. Xie, M. Bao, S. M. C. Bruekers, W. T. S. Huck, Collagen gels with different fibrillar microarchitectures elicit different cellular responses. *ACS Appl. Mater. Interfaces* **9**, 19630–19637 (2017).
53. T. Novak, B. Seelbinder, C. M. Twitchell, S. L. Voytik-Harbin, C. P. Neu, Dissociated and reconstituted cartilage microparticles in densified collagen induce local hMSC differentiation. *Adv. Funct. Mater.* **26**, 5427–5436 (2016).
54. L.-S. Wang, J. E. Chung, P. P. Chan, M. Kurisawa, Injectable biodegradable hydrogels with tunable mechanical properties for the stimulation of neurogenesic differentiation of human mesenchymal stem cells in 3D culture. *Biomaterials* **31**, 1148–1157 (2010).
55. S. D. Bolboacă, L. Jäntschi, Amino acids sequence analysis on collagen. *Bull. USAMV-CN* **64**, 311–316 (2007).
56. L. E. Bretscher, C. L. Jenkins, K. M. Taylor, M. L. DeRider, R. T. Raines, Conformational stability of collagen relies on a stereoelectronic effect. *J. Am. Chem. Soc.* **123**, 777–778 (2001).
57. B. Brodsky, G. Thiagarajan, B. Madhan, K. Kar, Triple-helical peptides: An approach to collagen conformation, stability, and self-association. *Biopolymers* **89**, 345–353 (2008).
58. B. Brodsky, A. V. Persikov, Molecular structure of the collagen triple helix. *Adv. Protein Chem.* **70**, 301–339 (2005).
59. J. A. Fallas, J. Dong, Y. J. Tao, J. D. Hartgerink, Structural insights into charge pair interactions in triple helical collagen-like proteins. *J. Biol. Chem.* **287**, 8039–8047 (2012).
60. A. L. Fidler, S. P. Boudko, A. Rokas, B. G. Hudson, The triple helix of collagens - an ancient protein structure that enabled animal multicellularity and tissue evolution. *J. Cell Sci.* **131**, jcs203950 (2018).
61. H.-P. Germann, E. Heidemann, A synthetic model of collagen: An experimental investigation of the triple-helix stability. *Biopolymers* **27**, 157–163 (1988).
62. I. Goldberga, R. Li, M. J. Duer, Collagen structure-function relationships from solid-state NMR spectroscopy. *Acc. Chem. Res.* **51**, 1621–1629 (2018).
63. C. L. Jenkins, R. T. Raines, Insights on the conformational stability of collagen. *Nat. Prod. Rep.* **19**, 49–59 (2002).
64. C. L. Jenkins, L. E. Bretscher, I. A. Guzei, R. T. Raines, Effect of 3-hydroxyproline residues on collagen stability. *J. Am. Chem. Soc.* **125**, 6422–6427 (2003).
65. K. Kar *et al.*, Aromatic interactions promote self-association of collagen triple-helical peptides to higher-order structures. *Biochemistry* **48**, 7959–7968 (2009).
66. M. V. Katti, R. Sami-Subbu, P. K. Ranjekar, V. S. Gupta, Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications. *Protein Sci.* **9**, 1203–1209 (2000).
67. K. T. Walker *et al.*, Non-linearity of the collagen triple helix in solution and implications for collagen function. *Biochem. J.* **474**, 2203–2217 (2017).
68. M. Sun *et al.*, Collagen V is a dominant regulator of collagen fibrillogenesis: Dysfunctional regulation of structure and function in a corneal-stroma-specific Col5a1-null mouse model. *J. Cell Sci.* **124**, 4096–4105 (2011).
69. M. D. Shoulders, J. A. Hodges, R. T. Raines, Reciprocity of steric and stereoelectronic effects in the collagen triple helix. *J. Am. Chem. Soc.* **128**, 8112–8113 (2006).
70. Y. Qiu *et al.*, Collagen Gly missense mutations: Effect of residue identity on collagen structure and integrin binding. *J. Struct. Biol.* **203**, 255–262 (2018).
71. A. V. Persikov, J. A. Ramshaw, B. Brodsky, Collagen model peptides: Sequence dependence of triple-helix stability. *Biopolymers* **55**, 436–450 (2000).
72. A. V. Persikov, Y. Xu, B. Brodsky, Equilibrium thermal transitions of collagen model peptides. *Protein Sci.* **13**, 893–902 (2004).
73. A. D. MacKerell *et al.*, All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
74. J. C. Phillips *et al.*, Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).
75. J. K. Rainey, M. C. Goh, An interactive triple-helical collagen builder. *Bioinformatics* **20**, 2458–2459 (2004).