SUMMARY

[0005a] In one aspect, the present disclosure provides a method of engineering one or more collagen molecules with at least one physical or chemical property, comprising:

(a) generating, using a machine learning model implemented on a computer system comprising one or more processors and system memory, a prediction that a set of target data comprising frequencies of amino acid residues in one or more target collagen sequences is associated with the at least one physical or chemical property, wherein the machine learning model was trained by:

(i) receiving a set of training data comprising frequencies of amino acid residues in a plurality of training collagen sequences and physical or chemical property data of the at least one physical or chemical property associated with the plurality of training collagen sequences,

wherein the frequencies of amino acid residues comprise intra-sequence variation of amino acid trimers in the plurality of training collagen sequences; and

(ii) training the machine learning model by fitting the machine learning model to the set of training data thereby generating a trained machine learning model, wherein the trained machine learning model is configured to receive as input amino acid frequency data of a test collagen sequence and predict at least one value of the at least one physical or chemical property associated with the test collagen sequence;

(b) subsequent to generating the prediction that the set of target data is associated with the at least one physical or chemical property, determining, by the computer system, one or more collagen sequences corresponding to the set of target data by identifying the one or more collagen sequences based at least in part on the frequencies of amino acid residues in the target data;

(c) producing one or more polynucleotides encoding the one or more collagen sequences; and

(d) expressing, on a protein production platform, the one or more polynucleotides to produce one or more collagen molecules comprising the one or more collagen sequences.

[0005b] In one aspect, the present disclosure provides a computer program product comprising a non-transitory machine readable medium storing program code that, when executed by one or more processors of a computer system, causes the computer system to

implement a method for engineering one or more collagen molecules with at least one physical or chemical property, said program code comprising:

(a) code for receiving a set of training data comprising frequencies of amino acid residues in a plurality of training collagen sequences and physical or chemical property data of at least one physical or chemical property associated with the plurality of training collagen sequences,

wherein the frequencies of amino acid residues comprise intra-sequence variation of amino acid trimers in the plurality of training collagen sequences; and

(b) code for training a machine learning model by fitting the machine learning model to the set of training data thereby generating a trained machine learning model, wherein the trained machine learning model is configured to receive as input amino acid frequency data of a test collagen sequence and predict at least one value of the at least one physical or chemical property associated with the test collagen sequence.

[0005c] In one aspect, the present disclosure provides a computer system, comprising:

one or more processors;

system memory; and

one or more computer-readable storage media having stored thereon computer-executable instructions that, when executed by the one or more processors, cause the computer system to implement a method for engineering one or more collagen molecules with at least one physical or chemical property, the one or more processors being configured to:

(a) receive a set of training data comprising frequencies of amino acid residues in a plurality of training collagen sequences and physical or chemical property data of at least one physical or chemical property associated with the plurality of training collagen sequences,

wherein the frequencies of amino acid residues comprise intra-sequence variation of amino acid trimers in the plurality of training collagen sequences; and

(b) train a machine learning model by fitting the machine learning model to the set of training data thereby generating a trained machine learning model, wherein the trained machine learning model is configured to receive as input amino acid frequency data of a test collagen sequence and predict at least one value of the at least one physical or chemical property associated with the test collagen sequence.

[0006] One example of the disclosure provides methods for engineering one or more collagen molecules. The method includes (a) obtaining, using a machine

learning model and by a computer system comprising one or more processors and system memory, a set of target data comprising frequencies of amino acid residues in one or more target collagen sequences, wherein the set of target data is predicted by the machine learning model to be associated with at least one physical or chemical property meeting a criterion, wherein the machine learning model was obtained by: (i) receiving a set of training data comprising frequencies of amino acid residues in a plurality of training collagen sequences and physical or chemical property data of the at least one physical or chemical property associated with the plurality of training collagen sequences; and (ii) training the machine learning model by fitting the machine learning model to the set of training data, wherein the trained machine learning model is configured to receive as input amino acid data of a test collagen sequence and predict at least one value of the at least one physical or chemical property associated with the test collagen sequence. The method also includes: (b) determining, by the computer system, one or more collagen sequences corresponding to the set of target data; (c) producing one or more polynucleotides encoding the one or more collagen sequences; and (d) expressing, on a protein production platform, the one or more polynucleotides to produce one or more collagen molecules comprising the one or more collagen sequences.

[0007] In some implementations, the frequencies of amino acid residues indicates intra-sequence variation of amino acid trimers in the plurality of collagen sequences. In some implementations, the frequencies of amino acid residues include: (a) a frequency for each of a plurality of different amino acids as residues at X positions of X-Y-Gly trimers in each training collagen sequence, and (b) a frequency for each of the different plurality of amino acids as residues at Y positions of the X-Y-Gly trimers in the training collagen sequence. In some implementations, the plurality of different amino acids includes 20 standard amino acids naturally occurring in organisms.

[0008] In some implementations, the plurality of amino acids further includes post-translational modifications of the 20 standard amino acids. In some

[0016] In some implementations, the plurality of training collagen sequences includes a plurality of gelatin sequences.

[0017] In some implementations, the at least one physical or chemical property is selected from a group consisting of: melting or gelling temperature, stiffness, elasticity, oxygen release rate, clarity, turbidity, ultraviolet blockage or absorption, viscosity, solubility, water content or hydration, resistance to protease, and ability to associate into fibrils.In some implementations, the at least one physical or chemical property includes two or more physical or chemical properties.

[0018] In some implementations, the one or more polynucleotides include recombinant polynucleotides. In some implementations, the one or more polynucleotides include synthesized polynucleotides.

[0019] In some implementations, the one or more collagen molecules produced in (d) include recombinant collagen molecules.

[0020] In some implementations, the method further includes manufacturing, using the one or more collagen molecules produced in (e), gelatin materials or collagen derivatives.

[0021] Another example of the disclosure provides a non-naturally occurring collagen polypeptide comprising: (a) an amino acid sequence of a secretion tag selected from the group consisting of DsbA, pelB, OmpA, TolB, MalE, lpp, TorA, and Hy1A; and (b) a plurality of X-Y-Gly trimers, wherein (i) amino acids at X positions of the X-Y-Gly trimers are selected from a group consisting of: alanine, cysteine, aspartic acid, glutamic acid, phenylalanine, glycine, histidine, isoleucine, lysine, leucine, methionine, asparagine, proline, pyrrolysine, glutamine, arginine, serine, threonine, selenocysteine, valine, tryptophan, tyrosine, and post-translational modifications therefrom, (ii) amino acids at Y positions of the X-Y-Gly trimers are selected from a group consisting of: alanine, cysteine, aspartic acid, glutamic acid, phenylalanine, glycine, histidine, isoleucine, lysine, leucine, methionine, asparagine, proline, pyrrolysine, glutamine, arginine, serine, threonine, selenocysteine, valine, tryptophan, tyrosine, and post-translational modifications therefrom, and (iii) the non-naturally occurring collagen polypeptide was predicted by a machine learning model to be associated with at least one physical or chemical property meeting a criterion.

[0029] In some implementations, the at least one physical or chemical property includes oxygen release rate.

[0030] In some implementations, the at least one physical or chemical property includes clarity.

[0031] In some implementations, the at least one physical or chemical property includes ultraviolet blockage or absorption.

[0032] In some implementations, the non-naturally occurring collagen polypeptide was produced by: (a) obtaining, using the machine learning model, a set of target data including frequencies of amino acid residues in one or more target collagen sequences, wherein the set of target data is predicted by the machine learning model to be associated with at least one physical or chemical property meeting a criterion; (b) determining one or more collagen sequences corresponding to the set of target data; and (c) producing the non-naturally occurring collagen polypeptide including the one or more collagen sequences.

[0033] An additional example of the disclosure provides a non-naturally occurring gelatin polypeptide including: (a) an amino acid sequence of a secretion tag selected from the group consisting of DsbA, pelB, OmpA, TolB, MalE, lpp, TorA, and Hy1A; and (b) a plurality of X-Y-Gly trimers, where (i) amino acids at X positions of the X-Y-Gly trimers are selected from a group consisting of: alanine, cysteine, aspartic acid, glutamic acid, phenylalanine, glycine, histidine, isoleucine, lysine, leucine, methionine, asparagine, proline, pyrrolysine, glutamine, arginine, serine, threonine, selenocysteine, valine, tryptophan, tyrosine, and post-translational modifications therefrom, (ii) amino acids at Y positions of the X-Y-Gly trimers are selected from a group consisting of: alanine, cysteine, aspartic acid, glutamic acid, phenylalanine, glycine, histidine, isoleucine, lysine, leucine, methionine, asparagine, proline, pyrrolysine, glutamine, arginine, serine, threonine, selenocysteine, valine, tryptophan, tyrosine, and post-translational modifications therefrom, and (iii) the non-naturally occurring gelatin polypeptide was predicted by a machine learning model to be associated with at least one physical or chemical property meeting a criterion.

[0034] Computer systems and computer program products for practicing the methods and making the compounds are also disclosed.

[0035] One example of the disclosure provides computer program product including a non-transitory machine readable medium storing program code that, when executed by one or more processors of a computer system, causes the computer system to implement a method for engineering one or more collagen molecules, said program code including: code for receiving a set of training data including frequencies of amino acid residues in a plurality of training collagen sequences and physical or chemical property data of at least one physical or chemical property associated with the plurality of training collagen sequences; and code for training a machine learning model by fitting the machine learning model to the set of training data, wherein the trained machine learning model is configured to receive as input amino acid data of a test collagen sequence and predict at least one value of the at least one physical or chemical property associated with the test collagen sequence.

[0036] In some implementations, the program code further includes: code for determining, using the machine learning model, a set of target data including frequencies of amino acid residues in one or more target collagen sequences, wherein the set of target data is predicted by the machine learning model to be associated with the at least one physical or chemical property meeting a criterion; and code for determining one or more collagen sequences corresponding to the set of target data.

[0037] Another example of the disclosure provides a computer system, including: one or more processors; system memory; and one or more computer-readable storage media having stored thereon computer-executable instructions that, when executed by the one or more processors, cause the computer system to implement a method for engineering one or more collagen molecules. The one or more processors are configured to: receive a set of training data including frequencies of amino acid residues in a plurality of training collagen sequences and physical or chemical property data of at least one physical or chemical property associated with the plurality of training collagen sequences; and train a machine learning model by fitting the machine learning model to the set of training data, wherein the trained machine learning model is configured to receive as input amino acid data of a test collagen sequence and predict at least one value of the at least one physical or chemical property associated with the test collagen sequence.

[0038] In some implementations, the one or more processors are further configured to: determine, using the machine learning model, a set of target data

CLAIMS

1. A method of engineering one or more collagen molecules with at least one physical or chemical property, comprising:

(a) generating, using a machine learning model implemented on a computer system comprising one or more processors and system memory, a prediction that a set of target data comprising frequencies of amino acid residues in one or more target collagen sequences is associated with the at least one physical or chemical property, wherein the machine learning model was trained by:

(i) receiving a set of training data comprising frequencies of amino acid residues in a plurality of training collagen sequences and physical or chemical property data of the at least one physical or chemical property associated with the plurality of training collagen sequences,

wherein the frequencies of amino acid residues comprise intra-sequence variation of amino acid trimers in the plurality of training collagen sequences; and

(ii) training the machine learning model by fitting the machine learning model to the set of training data thereby generating a trained machine learning model, wherein the trained machine learning model is configured to receive as input amino acid frequency data of a test collagen sequence and predict at least one value of the at least one physical or chemical property associated with the test collagen sequence;

(b) subsequent to generating the prediction that the set of target data is associated with the at least one physical or chemical property, determining, by the computer system, one or more collagen sequences corresponding to the set of target data by identifying the one or more collagen sequences based at least in part on the frequencies of amino acid residues in the target data;

(c) producing one or more polynucleotides encoding the one or more collagen sequences; and

(d) expressing, on a protein production platform, the one or more polynucleotides to produce one or more collagen molecules comprising the one or more collagen sequences.

2. The method of claim 1, wherein the frequencies of amino acid residues comprise: (a) a frequency for each of a plurality of different amino acids as residues at X positions of X-Y-Gly trimers in each training collagen sequence, and (b) a frequency for each of the plurality of different amino acids as residues at Y positions of the X-Y-Gly trimers in each training collagen sequence.

3. The method of claim 2, wherein the plurality of different amino acids comprises 20 standard amino acids naturally occurring in organisms and/or wherein the plurality of amino acids comprises post-translational modifications of the 20 standard amino acids.

4. The method of claim 2, wherein the plurality of amino acids consists of a subset of 20 standard amino acids and one or more post-translationally modified amino acids.

5. The method of any of claims 1-4, wherein the set of training data is generated using a main collagen domain with an uninterrupted $(X\text{-}Y\text{-}Gly)_n$ repeating sequence.

6. The method of any of claims 1-5, wherein the set of training data comprises lengths of the plurality of training collagen sequences.

7. The method of any of claims 1-6, wherein the frequencies of amino acid residues comprise: frequencies of amino acid residues in two or more regions of each training collagen sequence.

8. The method of any of claims 7, wherein the frequencies of amino acid residues comprise: (a) a frequency for each of a plurality of different amino acids at X positions of X-Y-Gly trimers in a first region of each training collagen sequence, (b) a frequency for each of a plurality of different amino acids at Y positions of X-Y-Gly trimers in the first region of each training collagen sequence, (c) a frequency for each of the plurality of different amino acids at the X positions of the X-Y-Gly trimers in a second region of each training collagen sequence, and (d) a frequency for each of the plurality of different amino acids at the Y positions of the X-Y-Gly trimers in the second region of each training collagen sequence.

9. The method of any of claims 1-8, wherein the machine learning model comprises a support vector machine.

10. The method of claim 9, wherein the support vector machine has a linear kernel or a nonlinear kernel.

11. The method of any of claims 1-10, wherein training the machine learning model comprises applying a linear support vector machine and a weight vector analysis to reduce dimensionality of a feature space, and/or wherein training the machine learning model comprises applying a principal component analysis to reduce dimensionality of feature space.

12. The method of any of claims 1-11, wherein the machine learning model comprises a random forest model, a neural network model, and/or a general linear model.

13. The method of any of claims 1-12, wherein the plurality of training collagen sequences comprises a plurality of collagen sequences, and/or a plurality of gelatin sequences.

14. The method of any of claims 1-13, wherein the at least one physical or chemical property is selected from a group consisting of: melting or gelling temperature, stiffness, elasticity, oxygen release rate, clarity, turbidity, ultraviolet blockage or absorption, viscosity, solubility, water content or hydration, resistance to protease, and ability to associate into fibrils; optionally, wherein the at least one physical or chemical property comprises two or more physical or chemical properties.

15. The method of any of claims 1-14, wherein the one or more polynucleotides comprise recombinant polynucleotides and/or synthesized polynucleotides.

16. The method of any of claims 1-15, wherein the one or more collagen molecules produced in (d) comprise recombinant collagen molecules.

17. The method of any of claims 1-16, further comprising manufacturing, using the one or more collagen molecules produced in (e), gelatin materials or collagen derivatives.

18. A computer program product comprising a non-transitory machine readable medium storing program code that, when executed by one or more processors of a computer system, causes

the computer system to implement a method for engineering one or more collagen molecules with at least one physical or chemical property, said program code comprising:

(a) code for receiving a set of training data comprising frequencies of amino acid residues in a plurality of training collagen sequences and physical or chemical property data of at least one physical or chemical property associated with the plurality of training collagen sequences,

wherein the frequencies of amino acid residues comprise intra-sequence variation of amino acid trimers in the plurality of training collagen sequences; and

(b) code for training a machine learning model by fitting the machine learning model to the set of training data thereby generating a trained machine learning model, wherein the trained machine learning model is configured to receive as input amino acid frequency data of a test collagen sequence and predict at least one value of the at least one physical or chemical property associated with the test collagen sequence.

19. The computer program product of claim 18, wherein said program code further comprising:

(c) code for determining, using the trained machine learning model, a set of target data comprising frequencies of amino acid residues in one or more target collagen sequences, wherein the set of target data is predicted by the trained machine learning model to be associated with the at least one physical or chemical property meeting a criterion; and

(d) code for determining one or more collagen sequences corresponding to the set of target data.

20. A computer system, comprising:

one or more processors;

system memory; and

one or more computer-readable storage media having stored thereon computer-executable instructions that, when executed by the one or more processors, cause the computer system to implement a method for engineering one or more collagen molecules with at least one physical or chemical property, the one or more processors being configured to:

(a) receive a set of training data comprising frequencies of amino acid residues in a plurality of training collagen sequences and physical or chemical property data of at least one physical or chemical property associated with the plurality of training collagen sequences,

wherein the frequencies of amino acid residues comprise intra-sequence variation of amino acid trimers in the plurality of training collagen sequences; and

(b) train a machine learning model by fitting the machine learning model to the set of training data thereby generating a trained machine learning model, wherein the trained machine learning model is configured to receive as input amino acid frequency data of a test collagen sequence and predict at least one value of the at least one physical or chemical property associated with the test collagen sequence.

21. The computer system of claim 20, wherein the one or more processors are further configured to:

(c) determine, using the trained machine learning model, a set of target data comprising frequencies of amino acid residues in one or more target collagen sequences, wherein the set of target data is predicted by the trained machine learning model to be associated with the at least one physical or chemical property meeting a criterion; and

(d) determine one or more collagen sequences corresponding to the set of target data.