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, the method comprising:

(a) generating, using a machine learning model implemented on a computer system comprising one or more processors and system memory, a prediction that indicates that a set of target data comprising relative 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 at least one physical or chemical property is selected from the group consisting of: 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, wherein the machine learning model is initially trained by:

(i) receiving a set of training data comprising (1) relative frequencies of amino acid residues in a plurality of training collagen sequences and (2) physical or chemical property data of the at least one physical or chemical property associated with the plurality of training collagen sequences, wherein a length of each of the plurality of training collagen sequences is at least about 100 amino acid residues; 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 (1) receive as input relative amino acid frequency data of a test collagen sequence and (2) 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, 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 relative 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, the program code comprising:

(a) code for receiving a set of training data comprising (1) relative frequencies of amino acid residues in a plurality of training collagen sequences and (2) physical or chemical property data of at least one physical or chemical property associated with the plurality of training collagen sequences,

wherein the at least one physical or chemical property is selected from the group consisting of: 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, and wherein a length of each of the plurality of training collagen sequences is at least about 100 amino acid residues;

(b) 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 (1) receive as input relative amino acid frequency data of a test collagen sequence and (2) predict at least one value of the at least one physical or chemical property associated with the test collagen sequence;

(c) code for determining, using the machine learning model, a prediction that indicates that a set of target data comprising relative frequencies of amino acid residues in one or more target collagen sequences is associated with the at least one physical or chemical property; and

(d) code for determining 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 relative frequencies of amino acid residues in the target data.

[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 (1) relative frequencies of amino acid residues in a plurality of training collagen sequences and (2) physical or chemical property data of at least one physical or chemical property associated with the plurality of training collagen sequences,

wherein the at least one physical or chemical property is selected from the group consisting of: 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, and wherein a length of each of the plurality of training collagen sequences is at least about 100 amino acid residues;

(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 (1) receive as input relative amino acid frequency data of a test collagen sequence and (2) predict at least one value of the at least one physical or chemical property associated with the test collagen sequence;

(c) determine, using the machine learning model, a prediction that indicates that a set of target data comprising relative frequencies of amino acid residues in one or more target collagen sequences is associated with the at least one physical or chemical property; and

(d) determine 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 relative frequencies of amino acid residues in the target data.

[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

CLAIMS

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

(a) generating, using a machine learning model implemented on a computer system comprising one or more processors and system memory, a prediction that indicates that a set of target data comprising relative 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 at least one physical or chemical property is selected from the group consisting of: 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, wherein the machine learning model is initially trained by:

(i) receiving a set of training data comprising (1) relative frequencies of amino acid residues in a plurality of training collagen sequences and (2) physical or chemical property data of the at least one physical or chemical property associated with the plurality of training collagen sequences, wherein a length of each of the plurality of training collagen sequences is at least about 100 amino acid residues; 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 (1) receive as input relative amino acid frequency data of a test collagen sequence and (2) 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, 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 relative 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 relative frequencies of amino acid residues comprise intra-sequence variation of amino acid trimers in the plurality of training collagen sequences.

3. The method of claim 2, wherein the relative 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.

4. The method of claim 3, wherein the plurality of different amino acids comprises 20 standard amino acids naturally occurring in organisms.

5. The method of claim 4, wherein the plurality of amino acids further comprises post-translational modifications of the 20 standard amino acids.

6. The method of claim 3, wherein the plurality of amino acids consists of a subset of 20 standard amino acids and one or more post-translationally modified amino acids of the subset.

7. The method of any of claims 1-6, wherein the set of training data is generated using a main collagen domain with an uninterrupted $(X\text{-}Y\text{-}Gly)_n$ repeating sequence.

8. The method of any of claims 1-7, wherein the set of training data comprises lengths of the plurality of training collagen sequences or fragments thereof.

9. The method of any of claims 1-8, wherein the relative frequencies of amino acid residues comprise: relative frequencies of amino acid residues in two or more regions of each training collagen sequence.

10. The method of any of claims 1-9, wherein the relative 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.

11. The method of any of claims 1-10, wherein the machine learning model comprises a support vector machine.

12. The method of claim 11, wherein the support vector machine has a linear kernel or a nonlinear kernel.

13. The method of any of claims 1-12, 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.

14. The method of any of claims 1-13, wherein the machine learning model comprises a random forest model, a neural network model, and/or a general linear model.

15. The method of any of claims 1-14, wherein the plurality of training collagen sequences comprises a plurality of collagen sequences, and/or a plurality of gelatin sequences.

16. The method of any of claims 1-15, wherein the at least one physical or chemical property comprises two or more physical or chemical properties.

17. The method of any of claims 1-16, wherein the one or more polynucleotides comprise recombinant polynucleotides and/or synthesized polynucleotides.

18. The method of any of claims 1-17, wherein the one or more collagen molecules produced in (d) comprise recombinant collagen molecules.

19. The method of any of claims 1-18, further comprising manufacturing, using the one or more collagen molecules produced in (e), gelatin materials or collagen derivatives.

20. 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, the program code comprising:

(a) code for receiving a set of training data comprising (1) relative frequencies of amino acid residues in a plurality of training collagen sequences and (2) physical or chemical property data of at least one physical or chemical property associated with the plurality of training collagen sequences, wherein the at least one physical or chemical property is selected from the group

consisting of: 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, and wherein a length of each of the plurality of training collagen sequences is at least about 100 amino acid residues;

(b) 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 (1) receive as input relative amino acid frequency data of a test collagen sequence and (2) predict at least one value of the at least one physical or chemical property associated with the test collagen sequence;

(c) code for determining, using the machine learning model, a prediction that indicates that a set of target data comprising relative frequencies of amino acid residues in one or more target collagen sequences is associated with the at least one physical or chemical property; and

(d) code for determining 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 relative frequencies of amino acid residues in the target data.

21. 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 (1) relative frequencies of amino acid residues in a plurality of training collagen sequences and (2) physical or chemical property data of at least one physical or chemical property associated with the plurality of training collagen sequences, wherein the at least one physical or chemical property is selected from the group consisting of: 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, and wherein a length of each of the plurality of training collagen sequences is at least about 100 amino acid residues;

(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 (1) receive as input relative amino acid frequency

data of a test collagen sequence and (2) predict at least one value of the at least one physical or chemical property associated with the test collagen sequence;

(c) determine, using the machine learning model, a prediction that indicates that a set of target data comprising relative frequencies of amino acid residues in one or more target collagen sequences is associated with the at least one physical or chemical property; and

(d) determine 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 relative frequencies of amino acid residues in the target data.