

CLAIMS

We claim:

1. A method for training a generative adversarial network (GAN), comprising:
 - a. generating, by a GAN generator, increasingly accurate positive simulated data until a GAN discriminator classifies the positive simulated data as positive;
 - b. presenting the positive simulated data, positive real data, and negative real data to a convolutional neural network (CNN), until the CNN classifies each type of data as positive or negative;
 - c. presenting the positive real data and the negative real data to the CNN to generate prediction scores; and
 - d. determining, based on the prediction scores, whether the GAN is trained or not trained, and when the GAN is not trained, repeating steps a-c until a determination is made, based on the prediction scores, that the GAN is trained.
2. The method of claim 1, wherein the positive simulated data, the positive real data, and the negative real data comprise biological data.
3. The method of claim 1, wherein the positive simulated data comprises positive simulated polypeptide-major histocompatibility complex class I (MHC-I) interaction data, the positive real data comprises positive real polypeptide-MHC-I interaction data, and the negative real data comprises negative real polypeptide-MHC-I interaction data.
4. The method of claim 3, wherein generating the increasingly accurate positive simulated polypeptide-MHC-I interaction data until the GAN discriminator classifies the positive simulated polypeptide-MHC-I interaction data as real comprises:

- e. generating, by the GAN generator according to a set of GAN parameters, a first simulated dataset comprising simulated positive polypeptide-MHC-I interactions for a MHC allele;
 - f. combining the first simulated dataset with the positive real polypeptide-MHC-I interactions for the MHC allele, and the negative real polypeptide-MHC-I interactions for the MHC allele to create a GAN training dataset;
 - g. determining, by a discriminator according to a decision boundary, whether a respective polypeptide-MHC-I interaction for the MHC allele in the GAN training dataset is simulated positive, real positive, or real negative;
 - h. adjusting, based on accuracy of the determination by the discriminator, one or more of the set of GAN parameters or the decision boundary; and
 - i. repeating steps e-h until a first stop criterion is satisfied.
5. The method of claim 4, wherein presenting the positive simulated polypeptide-MHC-I interaction data, the positive real polypeptide-MHC-I interaction data, and the negative real polypeptide-MHC-I interaction data to the convolutional neural network (CNN), until the CNN classifies respective polypeptide-MHC-I interaction data as positive or negative comprises:
- j. generating, by the GAN generator according to the set of GAN parameters, a second simulated dataset comprising simulated positive polypeptide-MHC-I interactions for the MHC allele;
 - k. combining the second simulated dataset, the positive real polypeptide-MHC-I interactions for the MHC allele, and the negative real polypeptide-MHC-I interactions for the MHC allele to create a CNN training dataset;
 - l. presenting the CNN training dataset to the convolutional neural network (CNN);

- m. classifying, by the CNN according to a set of CNN parameters, a respective polypeptide-MHC-I interaction for the MHC allele in the CNN training dataset as positive or negative;
 - n. adjusting, based on accuracy of the classification by the CNN, one or more of the set of CNN parameters; and
 - o. repeating steps l-n until a second stop criterion is satisfied.
- 6. The method of claim 5, wherein presenting the positive real polypeptide-MHC-I interaction data and the negative real polypeptide-MHC-I interaction data to the CNN to generate prediction scores comprises:
 - classifying, by the CNN according to the set of CNN parameters, a respective polypeptide-MHC-I interaction for the MHC allele as positive or negative.
- 7. The method of claim 6, wherein determining, based on the prediction scores, whether the GAN is trained comprises determining accuracy of the classification by the CNN, wherein when the accuracy of the classification satisfies a third stop criterion, outputting the GAN and the CNN.
- 8. The method of claim 6, wherein determining, based on the prediction scores, whether the GAN is trained comprises determining accuracy of the classification by the CNN, wherein when the accuracy of the classification does not satisfy a third stop criterion, returning to step a.
- 9. The method of claim 4, wherein the GAN parameters comprise one or more of allele type, allele length, generating category, model complexity, learning rate, or batch size.

10. The method of claim 9, wherein the allele type comprises one or more of HLA-A, HLA-B, HLA-C, or a subtype thereof.
11. The method of claim 9, wherein the allele length is from about 8 to about 12 amino acids.
12. The method of claim 11, wherein the allele length is from about 9 to about 11 amino acids.
13. The method of claim 3, further comprising:
 - presenting a dataset to the CNN, wherein the dataset comprises a plurality of candidate polypeptide-MHC-I interactions;
 - classifying, by the CNN, each of the plurality of candidate polypeptide-MHC-I interactions as a positive or a negative polypeptide-MHC-I interaction;
 - and
 - synthesizing the polypeptide from the candidate polypeptide-MHC-I interaction classified as a positive polypeptide-MHC-I interaction.
14. The polypeptide produced by the method of claim 13.
15. The method of claim 13, wherein the polypeptide is a tumor specific antigen.
16. The method of claim 13, wherein the polypeptide comprises an amino acid sequence that specifically binds to an MHC-I protein encoded by a selected MHC allele.
17. The method of claim 3, wherein the positive simulated polypeptide-MHC-I interaction data, the positive real polypeptide-MHC-I interaction data, and the

negative real polypeptide-MHC-I interaction data are associated with a selected allele.

18. The method of claim 17, wherein the selected allele is selected from a group consisting of A0201, A0202, A0203, B2703, B2705, and combinations thereof.
19. The method of claim 3, wherein generating the increasingly accurate positive simulated polypeptide-MHC-I interaction data until the GAN discriminator classifies the positive simulated polypeptide-MHC-I interaction data as positive comprises evaluating a gradient descent expression for the GAN generator.
20. The method of claim 3, wherein generating the increasingly accurate positive simulated polypeptide-MHC-I interaction data until the GAN discriminator classifies the positive simulated polypeptide-MHC-I interaction data as positive comprises:
 - iteratively executing the GAN discriminator in order to increase a likelihood of giving a high probability to positive real polypeptide-MHC-I interaction data, a low probability to the positive simulated polypeptide-MHC-I interaction data, and a low probability to the negative real polypeptide-MHC-I interaction data; and
 - iteratively executing the GAN generator in order to increase a probability of the positive simulated polypeptide-MHC-I interaction data being rated highly.
21. The method of claim 3, wherein presenting the positive simulated polypeptide-MHC-I interaction data, the positive real polypeptide-MHC-I interaction data, and the negative real polypeptide-MHC-I interaction data to the convolutional neural network (CNN), until the CNN classifies the respective polypeptide-MHC-I interaction data as positive or negative comprises:
 - performing a convolution procedure;
 - performing a Non Linearity (ReLU) procedure;

performing a Pooling or Sub Sampling procedure; and
performing a Classification (Fully Connected Layer) procedure.

22. The method of claim 1, wherein the GAN comprises a Deep Convolutional GAN (DCGAN).
23. The method of claim 8, wherein the first stop criterion comprises evaluating a mean squared error (MSE) function, the second stop criterion comprises evaluating a mean squared error (MSE) function, and the third stop criterion comprises evaluating an area under the curve (AUC) function.
24. The method of claim 3, wherein the prediction score is a probability of the positive real polypeptide-MHC-I interaction data being classified as positive polypeptide-MHC-I interaction data.
25. The method of claim 1, wherein determining, based on the prediction scores, whether the GAN is trained comprises comparing one or more of the prediction scores to a threshold.
26. The method of claim 1, further comprising outputting the GAN and the CNN.
27. An apparatus for training a generative adversarial network (GAN), comprising:
one or more processors; and
memory storing processor executable instructions that, when executed by the one or more processors, cause the apparatus to:
 - a. generate, by a GAN generator, increasingly accurate positive simulated data until a GAN discriminator classifies the positive simulated data as positive;

- b. present the positive simulated data, positive real data, and negative real data to a convolutional neural network (CNN), until the CNN classifies respective data as positive or negative;
 - c. present the positive real data and the negative real data to the CNN to generate prediction scores; and
 - d. determine, based on the prediction scores, whether the GAN is trained wherein when the GAN is not trained, repeating a-c until a determination is made, based on the prediction scores, that the GAN is trained.
28. The apparatus of claim 27, wherein the positive simulated data, the positive real data, and the negative real data comprise biological data.
29. The apparatus of claim 27, wherein the positive simulated data comprises positive simulated polypeptide-MHC-I interaction data, the positive real data comprises positive real polypeptide-MHC-I interaction data, and the negative real data comprises negative real polypeptide-MHC-I interaction data.
30. The apparatus of claim 29, wherein the processor executable instructions that, when executed by the one or more processors, cause the apparatus to generate the increasingly accurate positive simulated polypeptide-MHC-I interaction data until the GAN discriminator classifies the positive simulated polypeptide-MHC-I interaction data as positive further comprise processor executable instructions that, when executed by the one or more processors, cause the apparatus to:
- e. generate, according to a set of GAN parameters, a first simulated dataset comprising simulated positive polypeptide-MHC-I interactions for an MHC allele;

- f. combine the first simulated dataset the positive real polypeptide-MHC-I interactions for the MHC allele, and the negative real polypeptide-MHC-I interactions for the MHC allele to create a GAN training dataset;
 - g. receive information from a discriminator, wherein the discriminator is configured to determine, according to a decision boundary, whether a respective positive polypeptide-MHC-I interaction for the MHC allele in the GAN training dataset is positive or negative;
 - h. adjust, based on accuracy of the information from the discriminator, one or more of the set of GAN parameters or the decision boundary; and
 - i. repeat e-h until a first stop criterion is satisfied.
- 31. The apparatus of claim 30, wherein the processor executable instructions that, when executed by the one or more processors, cause the apparatus to present the positive simulated polypeptide-MHC-I interaction data, the positive real polypeptide-MHC-I interaction data, and the negative real polypeptide-MHC-I interaction data to a convolutional neural network (CNN), until the CNN classifies respective polypeptide-MHC-I interaction data as positive or negative further comprise processor executable instructions that, when executed by the one or more processors, cause the apparatus to:
 - j. generate, according to the set of GAN parameters, a second simulated dataset comprising simulated positive polypeptide-MHC-I interactions for the MHC allele;
 - k. combine the second simulated dataset, the positive real polypeptide-MHC-I interaction data for the MHC allele, and the negative real polypeptide-MHC-I interaction data for the MHC allele to create a CNN training dataset;
 - l. present the CNN training dataset to a convolutional neural network (CNN);
 - m. receive training information from the CNN, wherein the CNN is configured to determine the training information by classifying, according to a set of CNN

- parameters, a respective polypeptide-MHC-I interaction for the MHC allele in the CNN training dataset as positive or negative;
- n. adjust, based on accuracy of training information, one or more of the set of CNN parameters; and
 - o. repeat l-o until a second stop criterion is satisfied.
32. The apparatus of claim 31, wherein the processor executable instructions that, when executed by the one or more processors, cause the apparatus to present the positive real polypeptide-MHC-I interaction data and the negative real polypeptide-MHC-I interaction data to the CNN to generate prediction scores further comprise processor executable instructions that, when executed by the one or more processors, cause the apparatus to:
- classify, according to the set of CNN parameters, a respective polypeptide-MHC-I interaction for the MHC allele as positive or negative.
33. The apparatus of claim 32, wherein the processor executable instructions that, when executed by the one or more processors, cause the apparatus to determine, based on the prediction scores, whether the GAN is trained further comprise processor executable instructions that, when executed by the one or more processors, cause the apparatus to determine accuracy of the classification of the respective polypeptide-MHC-I interaction for the MHC allele as positive or negative, and when the accuracy of the classification satisfies a third stop criterion, output the GAN and the CNN.
34. The apparatus of claim 32, wherein the processor executable instructions that, when executed by the one or more processors, cause the apparatus to determine, based on the prediction scores, whether the GAN is trained further comprise processor executable instructions that, when executed by the one or more processors, cause the apparatus to determine accuracy of the classification of the respective polypeptide-

MHC-I interaction for the MHC allele as positive or negative, and when the accuracy of the classification does not satisfy a third stop criterion, return to step a.

35. The apparatus of claim 30, wherein the GAN parameters comprise one or more of allele type, allele length, generating category, model complexity, learning rate, or batch size.
36. The apparatus of claim 29, wherein the processor executable instructions, when executed by the one or more processors, further cause the apparatus to:
 - present a dataset to the CNN, wherein the dataset comprises a plurality of candidate polypeptide-MHC-I interactions, wherein the CNN is further configured to classify each of the plurality of candidate polypeptide-MHC-I interactions as a positive or a negative polypeptide-MHC-I interaction; and
 - synthesize the polypeptide from the candidate polypeptide-MHC-I interaction that the CNN classifies as a positive polypeptide-MHC-I interaction.
37. The apparatus of claim 29, wherein the positive simulated polypeptide-MHC-I interaction data, the positive real polypeptide-MHC-I interaction data, and the negative real polypeptide-MHC-I interaction data are associated with a selected allele.
38. The apparatus of claim 29, wherein the processor executable instructions that, when executed by the one or more processors, cause the apparatus to generate the increasingly accurate positive simulated polypeptide-MHC-I interaction data until the GAN discriminator classifies the positive simulated polypeptide-MHC-I interaction data as positive further comprise processor executable instructions that, when executed by the one or more processors, cause the apparatus to:

iteratively execute the GAN discriminator in order to increase a likelihood of giving a high probability to positive real polypeptide-MHC-I interaction data, a low probability to the positive simulated polypeptide-MHC-I interaction data, and a low probability to the negative simulated polypeptide-MHC-I interaction data; and

iteratively execute the GAN generator in order to increase a probability of the positive simulated polypeptide-MHC-I interaction data being rated highly.

39. The apparatus of claim 27, wherein the GAN comprises a Deep Convolutional GAN (DCGAN).
40. The apparatus of claim 33, wherein the first stop criterion comprises an evaluation of a mean squared error (MSE) function, the second stop criterion comprises an evaluation of a mean squared error (MSE) function, and the third stop criterion comprises an evaluation of an area under the curve (AUC) function.
41. The apparatus of claim 29, wherein the prediction score is a probability of the positive real polypeptide-MHC-I interaction data being classified as positive polypeptide-MHC-I interaction data.
42. The apparatus of claim 27, wherein the processor executable instructions that, when executed by the one or more processors, cause the apparatus to determine, based on the prediction scores, whether the GAN is trained further comprise processor executable instructions that, when executed by the one or more processors, cause the apparatus to compare one or more of the prediction scores to a threshold.
43. A non-transitory computer readable medium for training a generative adversarial network (GAN), the non-transitory computer readable medium storing processor

executable instructions that, when executed by one or more processors, causes the one or more processors to:

- a. generate, by a GAN generator, increasingly accurate positive simulated data until a GAN discriminator classifies the positive simulated data as positive;
- b. present the positive simulated data, positive real data, and negative real data to a convolutional neural network (CNN), until the CNN classifies respective data as positive or negative;
- c. present the positive real data and the negative real data to the CNN to generate prediction scores; and
- d. determine, based on the prediction scores, whether the GAN is trained wherein when the GAN is not trained, repeating a-c until a determination is made, based on the prediction scores, that the GAN is trained.

44. The non-transitory computer readable medium of claim 43, wherein the positive simulated data, the positive real data, and the negative real data comprise biological data.
45. The non-transitory computer readable medium of claim 43, wherein the positive simulated data comprises positive simulated polypeptide-MHC-I interaction data, the positive real data comprises positive real polypeptide-MHC-I interaction data, and the negative real data comprises negative real polypeptide-MHC-I interaction data.
46. The non-transitory computer readable medium of claim 45, wherein the processor executable instructions that, when executed by the one or more processors, cause the one or more processors to generate the increasingly accurate positive simulated polypeptide-MHC-I interaction data until the GAN discriminator classifies the positive simulated polypeptide-MHC-I interaction data as positive further cause the one or more processors to:

- e. generate, according to a set of GAN parameters, a first simulated dataset comprising simulated positive polypeptide-MHC-I interactions for an MHC allele;
 - f. combine the first simulated dataset with the positive real polypeptide-MHC-I interactions for the MHC allele, and the negative real polypeptide-MHC-I interactions for the MHC allele to create a GAN training dataset;
 - g. receive information from a discriminator, wherein the discriminator is configured to determine, according to a decision boundary, whether a respective positive polypeptide-MHC-I interaction for the MHC allele in the GAN training dataset is positive or negative;
 - h. adjust, based on accuracy of the information from the discriminator, one or more of the set of GAN parameters or the decision boundary; and
 - i. repeat e-h until a first stop criterion is satisfied.
47. The non-transitory computer readable medium of claim 46, wherein the processor executable instructions that, when executed by the one or more processors, cause the one or more processors to present the positive simulated polypeptide-MHC-I interaction data, the positive real polypeptide-MHC-I interaction data, and the negative real polypeptide-MHC-I interaction data to a convolutional neural network (CNN), until the CNN classifies respective polypeptide-MHC-I interaction data as positive or negative further comprise processor executable instructions that, when executed by the one or more processors, cause the one or more processors to:
- j. generate, according to the set of GAN parameters, a second simulated dataset comprising simulated positive polypeptide-MHC-I interactions for the MHC allele;
 - k. combine the second simulated dataset, the positive real polypeptide-MHC-I interaction data, and the negative real polypeptide-MHC-I interaction data for the MHC allele to create a CNN training dataset;

- l. present the CNN training dataset to a convolutional neural network (CNN);
 - m. receive training information from the CNN, wherein the CNN is configured to determine the training information by classifying, according to a set of CNN parameters, a respective polypeptide-MHC-I interaction for the MHC allele in the CNN training dataset as positive or negative;
 - n. adjust, based on accuracy of training information, one or more of the set of CNN parameters; and
 - o. repeat l-o until a second stop criterion is satisfied.
48. The non-transitory computer readable medium of claim 47, wherein the processor executable instructions that, when executed by the one or more processors, cause the one or more processors to present the positive real polypeptide-MHC-I interaction data and the negative real polypeptide-MHC-I interaction data to the CNN to generate prediction scores further comprise processor executable instructions that, when executed by the one or more processors, cause the one or more processors to:
 - p. present the CNN with the positive real polypeptide-MHC-I interaction data and the negative real polypeptide-MHC-I interaction data, wherein the CNN is further configured to classify, according to the set of CNN parameters, a respective polypeptide-MHC-I interaction for the MHC allele as positive or negative.
49. The non-transitory computer readable medium of claim 48, wherein the processor executable instructions that, when executed by the one or more processors, cause the one or more processors to determine, based on the prediction scores, whether the GAN is trained further comprise processor executable instructions that, when executed by the one or more processors, cause the one or more processors to determine accuracy of the classification of the respective polypeptide-MHC-I interaction for the MHC allele as positive or negative, and when the accuracy of the classification satisfies a third stop criterion, output the GAN and the CNN.

50. The non-transitory computer readable medium of claim 48, wherein the processor executable instructions that, when executed by the one or more processors, cause the one or more processors to determine, based on the prediction scores, whether the GAN is trained further comprise processor executable instructions that, when executed by the one or more processors, cause the one or more processors to determine accuracy of the classification of the respective polypeptide-MHC-I interaction for the MHC allele as positive or negative, and when the accuracy of the classification does not satisfy a third stop criterion, return to step a.
51. The non-transitory computer readable medium of claim 46, wherein the GAN parameters comprise one or more of allele type, allele length, generating category, model complexity, learning rate, or batch size.
52. The non-transitory computer readable medium of claim 45, wherein the processor executable instructions, when executed by the one or more processors, further cause the one or more processors to :
- present a dataset to the CNN, wherein the dataset comprises a plurality of candidate polypeptide-MHC-I interactions, wherein the CNN is further configured to classify each of the plurality of candidate polypeptide-MHC-I interactions as a positive or a negative polypeptide-MHC-I interaction; and
- synthesize the polypeptide from the candidate polypeptide-MHC-I interaction that the CNN classifies as a positive polypeptide-MHC-I interaction.
53. The non-transitory computer readable medium of claim 45, wherein the positive simulated polypeptide-MHC-I interaction data, the positive real polypeptide-MHC-I

interaction data, and the negative real polypeptide-MHC-I interaction data are associated with a selected allele.

54. The non-transitory computer readable medium of claim 45, wherein the processor executable instructions that, when executed by the one or more processors, cause the one or more processors to generate the increasingly accurate positive simulated polypeptide-MHC-I interaction data until the GAN discriminator classifies the positive simulated polypeptide-MHC-I interaction data as positive further comprise processor executable instructions that, when executed by the one or more processors, cause the one or more processors to:
- iteratively execute the GAN discriminator in order to increase a likelihood of giving a high probability to positive real polypeptide-MHC-I interaction data and a low probability the positive simulated polypeptide-MHC-I interaction data; and
 - iteratively execute the GAN generator in order to increase a probability of the positive simulated polypeptide-MHC-I interaction data being rated highly.
55. The non-transitory computer readable medium of claim 45, wherein the GAN comprises a Deep Convolutional GAN (DCGAN).
56. The non-transitory computer readable medium of claim 49, wherein the first stop criterion comprises an evaluation of a mean squared error (MSE) function, the second stop criterion comprises an evaluation of a mean squared error (MSE) function, and the third stop criterion comprises an evaluation of an area under the curve (AUC) function.
57. The non-transitory computer readable medium of claim 45, wherein the prediction score is a probability of the positive real polypeptide-MHC-I interaction data being classified as positive polypeptide-MHC-I interaction data.

58. The non-transitory computer readable medium of claim 45, wherein the processor executable instructions that, when executed by the one or more processors, cause the one or more processors to determine, based on the prediction scores, whether the GAN is trained further comprise processor executable instructions that, when executed by the one or more processors, cause the one or more processors to compare one or more of the prediction scores to a threshold.