CLAIMS

We claim:

1. A computer-implemented method for classifying binding of a peptide to a major histocompatibility complex class I (MHC-I) molecule, comprising:

   a. iteratively generating, by a computing device via a generative adversarial network (GAN) generator, increasingly accurate positive simulated polypeptide-MHC-I interaction data based on application of a noise vector to output an amino acid distribution matrix that mimics a polypeptide-MHC-I interaction, until a GAN discriminator classifies the positive simulated polypeptide-MHC-I interaction data as positive;

   b. training, by the computing device, a convolutional neural network (CNN) by presenting the positive simulated polypeptide-MHC-I interaction data, positive real polypeptide-MHC-I interaction data, and negative real polypeptide-MHC-I interaction data to the CNN, until the CNN classifies each type of data as positive or negative;

   c. presenting, by the computing device, the positive real polypeptide-MHC-I interaction data and the negative real polypeptide-MHC-I interaction data to the CNN to generate prediction scores; and

   d. determining, by the computing device, 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;

   presenting a dataset to the CNN, wherein the dataset comprises a plurality of candidate polypeptide-MHC-I interactions; and

   classifying, by the CNN, each of the plurality of candidate polypeptide-MHC-I interactions as a positive or a negative polypeptide-MHC-I interaction.

2. The computer-implemented method of claim 1, 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.

3. The computer-implemented method of claim 2, 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.

4. The computer-implemented method of claim 3, 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.

5. The computer-implemented method of claim 3, 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.

6. The computer-implemented method of claim 2, wherein the GAN parameters comprise one or more of allele type, allele length, generating category, model complexity, learning rate, or batch size.

7. The computer-implemented method of claim 6, wherein the allele type comprises one or more of HLA-A, HLA-B, HLA-C, or a subtype thereof.

8. The computer-implemented method of claim 1, further comprising:

synthesizing the polypeptide from the candidate polypeptide-MHC-I interaction classified as a positive polypeptide-MHC-I interaction.

9. A polypeptide produced by the method of claim 8.

10. The computer-implemented method of claim 8, wherein the polypeptide is a tumor specific antigen.

11. The computer-implemented method of claim 8, wherein the polypeptide comprises an amino acid sequence that specifically binds to an MHC-I protein encoded by a selected MHC allele.

12. The computer-implemented method of claim 1, 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.

13. The computer-implemented method of claim 1, 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.

14. The computer-implemented method of claim 1, 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.

15. The computer-implemented 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.

16. An apparatus configured for performing the method of any one of claims 1 to 8, 10 to 11 and 15.

17. A computer-readable medium (CRM) configured for performing the method of any one of claims 1 to 8, 10 to 11 and 15.