SUMMARY

[0004a] In one aspect, the present disclosure refers to a method comprising:

(a) processing amino acid information of a plurality of candidate peptide sequences using a machine learning HLA peptide presentation prediction model to generate a plurality of presentation predictions, wherein each candidate peptide sequence of the plurality of candidate peptide sequences is encoded by a genome or exome of a subject, or a pathogen or a virus in the subject, wherein the plurality of presentation predictions comprises an HLA presentation prediction for each of the plurality of candidate peptide sequences, wherein each HLA presentation prediction is indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject can present a given candidate peptide sequence of the plurality of candidate peptide sequences,

wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in training cells, wherein the training cells express an HLA allele with an affinity tag; and

~~(a) identifying a peptide sequence associated with cancer and comprising a cancer mutation, wherein identifying the peptide sequence comprises comparing DNA, RNA, or protein sequences from cancer cells of the subject to DNA, RNA, or protein sequences from normal cells of the subject;~~

(b) identifying, based at least on the plurality of presentation predictions, a peptide sequence of the plurality of peptide sequences as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject;

~~(b)~~ wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.07, wherein the PPV is calculated as a fraction of peptide sequences in the training data identified as being presented by a class II HLA allele of a cell in peptide sequences predicted to be presented by the class II HLA allele of the cell~~selecting a protein encoded by an HLA class II allele that is normally expressed by a cell of the subject, and predicted by a machine-learning HLA-peptide presentation prediction model to bind to the peptide; wherein the prediction model has a positive predictive value of at least 0.1 at a recall rate of at least 0.1%, from 0.1% 50% or at most 50% and wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the identified peptide sequence;~~

~~(c) contacting the identified peptide with the selected protein encoded by the HLA class II allele to verify whether the identified peptide competes with a placeholder peptide associated with the selected protein encoded by the HLA class II allele to displace the placeholder peptide with an IC50 value of less than 500 nM; and~~

~~(d) optionally, purifying the identified peptide~~.

[0004b] In another aspect, the present disclosure refers to a method comprising:

(a) processing amino acid information of a plurality of peptide sequences encoded by a genome or exome of a subject using a machine learning HLA peptide binding prediction model to generate a plurality of binding predictions, wherein the plurality of binding predictions comprises an HLA binding prediction for each of the plurality of candidate peptide sequences, each binding prediction indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject binds to a given candidate peptide sequence of the plurality of candidate peptide sequences,

wherein the machine learning HLA peptide binding prediction model is trained using training data comprising sequence information of sequences of peptides identified to bind to an HLA class II protein or a derivative thereof expressed in training cells, wherein the training cells express an HLA allele with an affinity tag; and

(b) identifying, based at least on the plurality of binding predictions, a peptide sequence of the plurality of peptide sequences that has a probability greater than a threshold binding prediction probability value of binding to at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject;

wherein the machine learning HLA peptide binding prediction model has a positive predictive value (PPV) of at least 0.1, wherein the PPV is calculated as a fraction of peptide sequences in the training data identified as being presented by a class II HLA allele of a cell in peptide sequences predicted to be presented by the class II allele of the cell~~of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, comprising:~~

~~obtaining, by a computer processor, a plurality of peptide sequences of the polypeptide sequence;~~

~~processing, by a computer processor, amino acid information of the plurality of peptide sequences using a machine-learning HLA-peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences, each presentation prediction indicative of a likelihood that one or more proteins encoded by a class I or II MHC allele of a cell of the subject can present an epitope sequence of a given peptide sequence of the plurality of peptide sequences, wherein the machine-learning HLA-peptide presentation prediction model is trained using training data comprising sequence information associated with the HLA protein expressed in cells;~~

~~determining or predicting that each of the plurality of peptide sequences of the polypeptide sequence would not be immunogenic to the subject based on the plurality of presentation predictions~~.

~~**[0004c]** In yet another aspect, the present disclosure refers to a method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, the method comprising:~~

~~(a) inputting amino acid information of peptide sequences of the polypeptide sequence, using a computer processor, into a machine-learning HLA-peptide presentation prediction model to generate a set of presentation predictions for the peptide sequences, each presentation prediction representing a probability that one or more proteins encoded by a class I or II MHC allele of a cell of the subject will present an epitope sequence of a given peptide sequence; wherein the machine-learning HLA-peptide~~

~~presentation prediction model comprises: a plurality of predictor variables identified at least based on training data, wherein the training data comprises: sequence information of sequences of peptides presented by a HLA protein expressed in cells and identified by mass spectrometry;~~

~~training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and the presentation likelihood generated as output based on the amino acid position information and the predictor variables; and~~

~~(b) determining or predicting that each of the peptide sequences of the polypeptide sequence would not be immunogenic to the subject based on the set of presentation predictions.~~

~~**[0004d]** In yet another aspect, the present disclosure refers to a use of a composition in the manufacture of a medicament for the treatment of an autoimmune disease or condition in a subject in need thereof, the composition comprising a regulatory T cell expressing a T cell receptor (TCR) that binds to a complex comprising: (i) an epitope of an expressed protein identified or predicted to be presented by a class I or II MHC of a cell of the subject, and (ii) the class I or II MHC, wherein the complex is targeted by a CD8 or CD4 T cell of the subject.~~

[0005] The methods and compositions described herein find uses in a wide range of applications. For example, the methods and compositions described herein can be used to identify immunogenic antigen

CLAIMS

1. A method comprising:

(a) processing amino acid information of a plurality of candidate peptide sequences using a machine learning HLA peptide presentation prediction model to generate a plurality of presentation predictions, wherein each candidate peptide sequence of the plurality of candidate peptide sequences is encoded by a genome or exome of a subject, or a pathogen or a virus in the subject, wherein the plurality of presentation predictions comprises an HLA presentation prediction for each of the plurality of candidate peptide sequences, wherein each HLA presentation prediction is indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject can present a given candidate peptide sequence of the plurality of candidate peptide sequences,

wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in training cells, wherein the training cells express an HLA allele with an affinity tag; and

(b) identifying, based at least on the plurality of presentation predictions, a peptide sequence of the plurality of peptide sequences as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject;

wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.07, wherein the PPV is calculated as a fraction of peptide sequences in the training data identified as being presented by a class II HLA allele of a cell in peptide sequences predicted to be presented by the class II HLA allele of the cell ~~according to a presentation PPV determination method~~.

2. A method comprising:

(a) processing amino acid information of a plurality of peptide sequences encoded by a genome or exome of a subject using a machine learning HLA peptide binding prediction model to generate a plurality of binding predictions, wherein the plurality of binding predictions comprises an HLA binding prediction for each of the plurality of candidate peptide sequences, each binding prediction indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject binds to a given candidate peptide sequence of the plurality of candidate peptide sequences,

wherein the machine learning HLA peptide binding prediction model is trained using training data comprising sequence information of sequences of peptides identified to

bind to an HLA class II protein or ~~an HLA class II protein analog~~a derivative thereof expressed in training cells, wherein the training cells express an HLA allele with an affinity tag; and

(b) identifying, based at least on the plurality of binding predictions, a peptide sequence of the plurality of peptide sequences that has a probability greater than a threshold binding prediction probability value of binding to at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject;

wherein the machine learning HLA peptide binding prediction model has a positive predictive value (PPV) of at least 0.1, wherein the PPV is calculated as a fraction of peptide sequences in the training data identified as being presented by a class II HLA allele of a cell in peptide sequences predicted to be presented by the class II allele of the cell ~~according to a binding PPV determination method~~.

3. The method of claim 2, wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in the training cells.

4. The method of ~~any one of the preceding claims~~claim 1 or 2, wherein the method comprises ranking, based on the presentation predictions, at least two peptides identified as being presented by at least one of the one or more proteins encoded by ~~a~~ the class II HLA allele of ~~a~~ the cell of the subject.

5. The method of ~~any one of the preceding claims~~claim 1 or 2, wherein the method comprises selecting one or more peptides of the two or more ranked peptides.

6. The method ~~of any one of the preceding claims~~claim 1 or 2, wherein the method comprises selecting one or more peptides of the plurality that were identified as being presented by at least one of the one or more proteins encoded by ~~a~~ the class II HLA allele of ~~a~~ the cell of the subject.

7. The method of ~~any one of the preceding claims~~claim 1 or 2, wherein the method comprises selecting one or more peptides of two or more peptides ranked based on the presentation predictions.

8. The method of ~~any one of the preceding claims~~claim 1, wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.07 when amino acid information of a plurality of test peptide sequences are processed to generate a plurality of test presentation predictions, each test presentation prediction indicative of a likelihood that the one or more proteins encoded by ~~a~~ the class II HLA allele of ~~a~~ the cell of the subject can present a given test peptide sequence of the plurality of test peptide sequences, wherein the plurality of test peptide sequences comprises at least 500 test peptide sequences comprising (i) at least one hit peptide

sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells and (ii) at least 499 decoy peptide sequences contained within a protein encoded by a genome of an organism, wherein the organism and the subject are the same species, wherein the plurality of test peptide sequences comprises a ratio of 1:499 of the at least one hit peptide sequence to the at least 499 decoy peptide sequences and a top percentage of the plurality of test peptide sequences are predicted to be presented by the HLA protein expressed in cells by the machine learning HLA peptide presentation prediction model.

9. The method of ~~any one of the preceding claims~~claim 2, wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.1 when amino acid information of a plurality of test peptide sequences are processed to generate a plurality of test binding predictions, each test binding prediction indicative of a likelihood that the one or more proteins encoded by ~~a~~ the class II HLA allele of ~~a~~ the cell of the subject binds to a given test peptide sequence of the plurality of test peptide sequences, wherein the plurality of test peptide sequences comprises at least 20 test peptide sequences comprising (i) at least one hit peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells and (ii) at least 19 decoy peptide sequences contained within a protein comprising at least one peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells, such as a single HLA protein expressed in cells (e.g., mono-allelic cells), wherein the plurality of test peptide sequences comprises a ratio of 1:19 of the at least one hit peptide sequence to the at least 19 decoy peptide sequences and a top percentage of the plurality of test peptide sequences are predicted to bind to the HLA protein expressed in cells by the machine learning HLA peptide presentation prediction model.

10. The method of ~~any one of the preceding claims~~claims 8 or 9, wherein no amino acid sequence overlap exist among the at least one hit peptide sequence and the decoy peptide sequences.

11. The method of any one of claims 1-10, wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.2, 0.21, 0.22, 0.23, 0.24, 0.25, 0.26, 0.27, 0.28, 0.29, 0.3, 0.31, 0.32, 0.33, 0.34, 0.35, 0.36, 0.37, 0.38, 0.39, 0.4, 0.41, 0.42, 0.43, 0.44, 0.45, 0.46, 0.47, 0.48, 0.49, 0.5, 0.51, 0.52, 0.53, 0.54, 0.55, 0.56, 0.57, 0.58, 0.59, 0.6, 0.61, 0.62, 0.63, 0.64, 0.65, 0.66, 0.67, 0.68, 0.69, 0.7, 0.71, 0.72, 0.73, 0.74, 0.75, 0.76, 0.77, 0.78, 0.79, 0.8, 0.81, 0.82, 0.83, 0.84, 0.85, 0.86, 0.87, 0.88, 0.89, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98 or 0.99.

12. The method of any one of claims ~~1 and 3-11~~8-11, wherein the at least one hit peptide sequence comprises at least 5, 10, 20, 50 or 100 hit peptide sequences.

13. The method of any one of claims ~~1 and 3-12~~8 and 10-12, wherein the at least 499 decoy peptide sequences comprises at least 2500, 5000, 10000, 25000, 50000 or 100000 decoy peptide sequences.

14. The method of any one of claims ~~1 and 3-13~~8 and 10-13, wherein the at least 500 test peptide sequences comprises at least 2500, 5000, 10000, 25000, 50000 or 100000 test peptide sequences.

15. The method of any one of claims ~~1 and 3-14~~8-14, wherein the top percentage of the plurality of test peptide sequences is a top 0.2%, 0.3%, 0.4%, 0.5%, 0.6%, 0.7%, 0.8%, 0.9%, 1.00% or 2.00%.

~~16. The method of any one of claims 2-11, wherein the at least one hit peptide sequence comprises at least 5, 10, 20, 50 or 100 hit peptide sequences.~~

~~17.~~16. The method of any one of claims ~~2-11 and 16~~9-12, wherein the at least 19 decoy peptide sequences comprises at least 500, 1000, 2000, 5000 or 10000 decoy peptide sequences.

~~18.~~17. The method of any one of claims ~~2-11, 16 and 17~~9-12 and 16, wherein the at least 20 test peptide sequences comprises at least wherein the at least 500 test peptide sequences comprises at least 500, 1000, 2000, 5000 or 10000 test peptide sequences test peptide sequences.

~~19. The method of any one of claims 2-11 and 16-18, wherein the top percentage is a top 5%, 6%, 7%, 8%, 9%, 10%, 15%, or 20%.~~

~~20.~~18. The method of any one of the preceding claims, wherein the PPV is more than the PPV of column 2 of Table 11 for the protein encoded by the corresponding HLA allele of Table 11.

~~21.~~19. The method of any one of the preceding claims, wherein the PPV is more than the PPV of column 2 of Table 16 for the protein encoded by ~~an~~ the HLA class II allele.

~~22.~~20. The method of any one of the preceding claims, wherein the subject is a single subject.

~~23.~~21. The method of any one of the preceding claims, wherein the training cells are cells expressing a single protein encoded by ~~a~~ the class II HLA allele of ~~a~~ the cell of the subject.

~~24.~~22. The method of any one of claims 1-~~23~~21, wherein the training cells are monoallelic HLA cells~~, or cells expressing an HLA allele with an affinity tag~~.

~~25.~~23. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is associated with a cancer.

~~26.~~24. The method of any one of the preceding claims, wherein at least one peptide sequence of the plurality of peptide sequences is overexpressed by a cancer cell of the subject.

~~27.~~25. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is overexpressed by a cancer cell of the subject.

~~28.~~26. The method of any one of the preceding claims, wherein at least one peptide sequence of the plurality of peptide sequences is a cancer cell-specific peptide.

~~29.~~27. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is a cancer cell-specific peptide.

~~30.~~28. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is expressed by a cancer cell of the subject.

~~31.~~29. The method of any one of the preceding claims, wherein at least one peptide sequence of the plurality of peptide sequences is not encoded by a non-cancer cell of the subject.

~~32.~~30. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is not encoded by a non-cancer cell of the subject.

~~33.~~31. The method of any one of the preceding claims, wherein at least one peptide sequence of the plurality of peptide sequences is not expressed by a non-cancer cell of the subject.

~~34.~~32. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is not expressed by a non-cancer cell of the subject.

~~35.~~33. The method of any one of the preceding claims, wherein the method comprises obtaining the plurality of peptide sequences of the subject.

~~36.~~34. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject.

~~37.~~35. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject that encodes the plurality of peptide sequences encoded by a genome or exome of a subject.

~~38.~~36. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject that encodes the plurality of peptide sequences encoded by a genome or exome of a subject by a computer processor.

~~39.~~37. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject by genomic or exome sequencing.

~~40.~~38. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject by whole genome sequencing or whole exome sequencing.

~~41.~~39. The method of any one of the preceding claims, wherein the method further comprises formulating a composition comprising one or more of the selected sub-set of peptide sequences.

~~42.~~40. The method of any one of the preceding claims, wherein the HLA class II protein comprises an HLA-DR, HLA-DQ, or an HLA-DP protein.

~~43.~~41. The method of any one of the preceding claims, wherein the HLA class II protein is selected from the group consisting of: HLA-DPB1*01:01/HLA-DPA1*01:03, HLA-DPB1*02:01/HLA-

DPA1*01:03, HLA-DPB1*03:01/HLA-DPA1*01:03, HLA-DPB1*04:01/HLA-DPA1*01:03, HLA-DPB1*04:02/HLA-DPA1*01:03, HLA-DPB1*06:01/HLA-DPA1*01:03,HLA-DQB1*02:01/HLA-DQA1*05:01, HLA-DQB1*02:02/HLA-DQA1*02:01, HLA-DQB1*06:02/HLA-DQA1*01:02, HLA-DQB1*06:04/HLA-DQA1*01:02, HLA-DRB1*01:01, HLA-DRB1*01:02, HLA-DRB1*03:01, HLA-DRB1*03:02, HLA-DRB1*04:01, HLA-DRB1*04:02, HLA-DRB1*04:03, HLA-DRB1*04:04, HLA-DRB1*04:05, HLA-DRB1*04:07, HLA-DRB1*07:01, HLA-DRB1*08:01, HLA-DRB1*08:02, HLA-DRB1*08:03, HLA-DRB1*08:04, HLA-DRB1*09:01, HLA-DRB1*10:01, HLA-DRB1*11:01, HLA-DRB1*11:02, HLA-DRB1*11:04, HLA-DRB1*12:01, HLA-DRB1*12:02, HLA-DRB1*13:01, HLA-DRB1*13:02, HLA-DRB1*13:03, HLA-DRB1*14:01, HLA-DRB1*15:01, HLA-DRB1*15:02, HLA-DRB1*15:03, HLA-DRB1*16:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB3*03:01, HLA-DRB4*01:01 and HLA-DRB5*01:01; an HLA-DR protein selected from the group consisting of: DPA*01:03/DPB*04:01, DRB1*01:01, DRB1*01:02, DRB1*03:01, DRB1*04:01, DRB1*04:02, DRB1*04:04, DRB1*04:05, DRB1*07:01, DRB1*08:01, DRB1*08:02, DRB1*08:03, DRB1*09:01, DRB1*11:01, DRB1*11:02, DRB1*11:04, DRB1*12:01, DRB1*13:01, DRB1*13:02, DRB1*13:03, DRB1*14:01, DRB1*15:01, DRB1*15:02, DRB1*15:03, DRB1*16:02, DRB3*01:01, DRB3*02:01, DRB3*02:02, DRB3*03:01, DRB4*01:01, DRB4*01:03 and DRB5*01:01; an HLA-DP protein selected from the group consisting of: DPB1*01:01, DPB1*02:01, DPB1*02:02, DPB1*03:01, DPB1*04:01, DPB1*04:02, DPB1*05:01, DPB1*06:01, DPB1*11:01, DPB1*13:01, DPB1*17:01; or an HLA-DQ protein complex selected from the group consisting of: A1*01:01+B1*05:01, A1*01:02+B1*06:02, A1*01:02+B1*06:04, A1*01:03+B1*06:03, A1*02:01+B1*02:02, A1*02:01+B1*03:03, A1*03:01+B1*03:02, A1*03:03+B1*03:01, A1*05:01+B1*02:01 and A1*05:05+B1*03:01.

~~44.~~42. The method of any one of the preceding claims, wherein the peptides presented by the HLA protein have a length of from 15-40 amino acids.

~~45.~~43. The method of ~~any one of the preceding claims~~claim 39, wherein the composition ~~personalized cancer therapy~~ further comprises an adjuvant.

~~46.~~44. The method of ~~any one of the preceding claims~~claim 39, wherein the ~~personalized cancer therapy~~composition further comprises an immune checkpoint inhibitor.

~~47. A method for identifying HLA class II specific peptides for immunotherapy for a subject, comprising:~~

(a) obtaining, by a computer processor, a candidate peptide comprising an epitope, and a plurality of peptide sequences, each comprising the epitope;

(b) processing, by a computer processor, amino acid information of the plurality of peptide sequences using a machine learning HLA peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences to an immune cell, each presentation prediction indicative of a likelihood that one or more proteins encoded by an HLA class II allele can present a given peptide sequence of the plurality of peptide sequences, wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of peptides presented by an HLA protein expressed in cells and identified by mass spectrometry;

(c) selecting a protein from the one or more proteins encoded by the HLA class II allele of a cell of the subject, predicted to bind to the candidate peptide by the machine learning HLA peptide presentation prediction model, wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the candidate peptide to an immune cell;

(d) contacting the candidate peptide with the selected protein such that the candidate peptide competes with a placeholder peptide associated with the selected protein; and

(e) identifying the candidate peptide as a peptide for immunotherapy specific for the selected protein based on whether the candidate peptide displaces the placeholder peptide.

48. A method for assaying immunogenicity of a MHC class II binding peptide, comprising:

(a) selecting a protein encoded by an HLA class II allele predicted by a machine learning HLA peptide presentation prediction model to bind to the MHC class II binding peptide, wherein the machine learning HLA peptide presentation prediction model is configured to generate a presentation prediction for a given peptide sequence, the presentation prediction indicative of a likelihood that one or more proteins encoded by the HLA class II allele can present the given peptide sequence, and wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the MHC class II binding peptide;

(b) contacting the peptide with the selected protein such that the peptide competes with a placeholder peptide associated with the selected protein, and displaces the placeholder peptide, thereby forming a complex comprising the HLA class II protein and the MHC class II binding peptide;

(c) contacting the complex with a CD4+ T cell, and

(d) assaying for one or more of activation parameters of the CD4+T cell, selected from the group consisting of: induction of a cytokine, induction of a chemokine, and expression of a cell surface marker.

49. A method comprising:

(a) identifying a peptide sequence associated with cancer and comprising a cancer mutation, wherein identifying the peptide sequence comprises comparing DNA, RNA, or protein sequences from cancer cells of the subject to DNA, RNA, or protein sequences from normal cells of the subject;

(b) selecting a protein encoded by an HLA class II allele that is normally expressed by a cell of the subject, and predicted by a machine learning HLA peptide presentation prediction model to bind to the peptide; wherein the prediction model has a positive predictive value of at least 0.1 at a recall rate of at least 0.1%, from 0.1% 50% or at most 50% and wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the identified peptide sequence;

(c) contacting the identified peptide with the selected protein encoded by the HLA class II allele to verify whether the identified peptide competes with a placeholder peptide associated with the selected protein encoded by the HLA class II allele to displace the placeholder peptide with an IC50 value of less than 500 nM; and

(d) optionally, purifying the identified peptide.

50. A method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, comprising:

(a) obtaining, by a computer processor, a plurality of peptide sequences of the polypeptide sequence;

(b) processing, by a computer processor, amino acid information of the plurality of peptide sequences using a machine learning HLA peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences, each presentation prediction indicative of a likelihood that one or more proteins encoded by a class I or II MHC allele of a cell of the subject can present an epitope sequence of a given peptide sequence of the plurality of peptide sequences, wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information associated with the HLA protein expressed in cells;

(c) determining or predicting that each of the plurality of peptide sequences of the polypeptide sequence would not be immunogenic to the subject based on the plurality of presentation predictions.

51. A method for manufacturing HLA class II tetramers by conjugation of four individual HLA protein alpha1 and beta1 heterodimers, the method comprising:

(a) expressing in a eukaryotic cell, a vector comprising a nucleic acid sequence encoding an alpha chain and a beta chain of HLA protein, a secretion signal, a biotinylation motif and at least one tag for identification or for purification, such that each HLA protein alpha-1 and beta1 heterodimers is secreted in dimerized state, wherein the heterodimer is associated with a placeholder peptide,

(b) purifying the secreted heterodimer from cell medium,

(c) validating the peptide binding activity using peptide exchange assay,

(d) adding streptavidin thereby conjugating heterodimers into tetramers,

(e) purifying the tetramers and having a yield of greater than 1 mg/L.

52. An HLA class II tetramer or multimer comprising either HLA DR, or HLA DP, or HLA DQ heterodimers, each heterodimer comprising an alpha and a beta chain, wherein the heterodimer is purified and present at a concentration of greater than 1 mg/L.

53. The HLA class II tetramer or multimer of claim 52, wherein the HLA protein is an HLA class II protein selected from the group consisting of: HLA-DPB1*01:01/HLA-DPA1*01:03, HLA-DPB1*02:01/HLA-DPA1*01:03, HLA-DPB1*03:01/HLA-DPA1*01:03, HLA-DPB1*04:01/HLA-DPA1*01:03, HLA-DPB1*04:02/HLA-DPA1*01:03, HLA-DPB1*06:01/HLA-DPA1*01:03,HLA-DQB1*02:01/HLA-DQA1*05:01,HLA-DQB1*02:02/HLA-DQA1*02:01, HLA-DQB1*06:02/HLA-DQA1*01:02,HLA-DQB1*06:04/HLA-DQA1*01:02, HLA-DRB1*01:01, HLA-DRB1*01:02, HLA-DRB1*03:01, HLA-DRB1*03:02, HLA-DRB1*04:01, HLA-DRB1*04:02, HLA-DRB1*04:03, HLA-DRB1*04:04, HLA-DRB1*04:05, HLA-DRB1*04:07, HLA-DRB1*07:01, HLA-DRB1*08:01, HLA-DRB1*08:02, HLA-DRB1*08:03, HLA-DRB1*08:04, HLA-DRB1*09:01, HLA-DRB1*10:01, HLA-DRB1*11:01, HLA-DRB1*11:02, HLA-DRB1*11:04, HLA-DRB1*12:01, HLA-DRB1*12:02, HLA-DRB1*13:01, HLA-DRB1*13:02, HLA-DRB1*13:03, HLA-DRB1*14:01, HLA-DRB1*15:01, HLA-DRB1*15:02, HLA-DRB1*15:03, HLA-DRB1*16:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB3*03:01, HLA-DRB4*01:01, and HLA-DRB5*01:01.

54. A vector comprising a nucleic acid sequence encoding an alpha chain and a beta chain of HLA protein of any one of the preceding claims, a secretion signal, a biotinylation motif and at least one tag for identification or for purification, such that each HLA protein alpha 1 and beta1 heterodimers is secreted in dimerized state, optionally, wherein the secreted heterodimer is associated with a placeholder peptide.

55. A cell, comprising the vector of claim 54.

56. A method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, comprising:

(a) obtaining, by a computer processor, a plurality of peptide sequences of the polypeptide sequence;

(b) processing, by a computer processor, amino acid information of the plurality of peptide sequences using a machine learning HLA peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences, each presentation prediction indicative of a likelihood that one or more proteins encoded by a class I or II MHC allele of a cell of the subject can present an epitope sequence of a given peptide sequence of the plurality of peptide sequences, wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of peptides presented by a HLA protein expressed in cells and identified by mass spectrometry; and

(c) determining or predicting that at least one of the plurality of peptide sequences of the polypeptide sequence would be immunogenic to the subject based on the plurality of presentation predictions.

57. A method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, the method comprising:

(a) inputting amino acid information of peptide sequences of the polypeptide sequence, using a computer processor, into a machine learning HLA peptide presentation prediction model to generate a set of presentation predictions for the peptide sequences, each presentation prediction representing a probability that one or more proteins encoded by a class I or II MHC allele of a cell of the subject will present an epitope sequence of a given peptide sequence; wherein the machine learning HLA peptide presentation prediction model comprises: a plurality of predictor variables identified at least based on training data, wherein the training data comprises: sequence information of sequences of peptides presented by a HLA protein expressed in cells and identified by mass spectrometry;

training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and the presentation likelihood generated as output based on the amino acid position information and the predictor variables; and

(b) determining or predicting that each of the peptide sequences of the polypeptide sequence would not be immunogenic to the subject based on the set of presentation predictions.

58. A method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, the method comprising:

(a) inputting amino acid information of peptide sequences of the polypeptide sequence, using a computer processor, into a machine-learning HLA-peptide presentation prediction model to generate a set of presentation predictions for the peptide sequences, each presentation prediction representing a probability that one or more proteins encoded by a class I or II MHC allele of a cell of the subject will present an epitope sequence of a given peptide sequence; wherein the machine learning HLA peptide presentation prediction model comprises: a plurality of predictor variables identified at least based on training data; wherein the training data comprises: sequence information of sequences of peptides presented by a HLA protein expressed in cells and identified by mass spectrometry; training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and the presentation likelihood generated as output based on the amino acid position information and the predictor variables;

(b) determining or predicting that at least one of the peptide sequences of the polypeptide sequence would be immunogenic to the subject based on the set of presentation predictions.

59. Use of a composition in the manufacture of a medicament for the treatment of an autoimmune disease or condition in a subject in need thereof, the composition comprising a regulatory T cell expressing a T cell receptor (TCR) that binds to a complex comprising: (i) an epitope of an expressed protein identified or predicted to be presented by a class I or II MHC of a cell of the subject, and (ii) the class I or II MHC, wherein the complex is targeted by a CD8 or CD4 T cell of the subject.