

CLAIMS

What is claimed is:

1. A method of identifying one or more antigenic peptides comprising:
 - a) obtaining one or more tissue samples,
 - b) acquiring mass spectrometry spectra for one or more antigenic peptides;
 - c) comparing the mass spectrometry data spectra to peptide theoretical spectra located in one or more public or non-public databases,
 - d) generating a peptide spectrum match (PSM) of the one or more antigenic peptides,
 - e) producing a matched spectral library or database of antigenic peptides based on steps (a) -- (d),
 - f) using a deep learning algorithm to train at least 80% of the peptide data located in the database or spectral library and testing the balance of the peptide data located in the database or library thereby producing a peptide prediction model to generate predicted peptide spectrum;
 - g) using the prediction model to identify one or more antigenic peptides.
2. The method of claim 1, wherein the mass spectrometry comprises tandem mass spectrometry (MS/MS).
3. The method of any one of claim 1 or 2, wherein the library or database comprises over about 70%, over about 80%, over about 85%, over about 90%, over about 95%, or 100% antigenic peptide data.
4. The method of any one of claims 1–3, wherein the library or database comprises less than about 30%, less than about 25%, less than about 20%, less than about 15%, less than about 10%, or less than about 5% tryptic peptides data.
5. The method of any one of claims 1–4, wherein the one or more antigenic peptides identified by the predicted spectra have an identification correlation within about 2% to about 15% relative to the actual technical variation of the experimentally determined spectra.
6. The method of any one of claims 1–4, wherein the prediction peptide performance score is greater than about 0.95.
7. The method of any one of claims 1–4, wherein the prediction peptide performance score is from about 0.92 to about 0.98.
8. The method of any one of claims 1–7, wherein the peptide spectrum match (PSM) have a false discovery rate (FDR) of less than 0.05.
9. The method of any one of claims 1–8, wherein antigenic peptides are identified with greater accuracy than tryptic peptides.

10. The method of any one of claims 1–9, wherein the one or more identified antigenic peptides exhibit a peptide performance score that is closer to the measured technical variation as compared to analyzing the same one or more peptides with ProteomeTools.
11. The method of any one of claims 1–10, wherein the antigenic peptides are 8 to 11 amino acid or 8 to 9 amino acids in length.
12. The method of any one of claims 1–11, wherein the one or more antigenic peptides identified are overexpressed or presented in one or more specific cancer tissues.
13. The method of any one of claims 1–11, wherein the tissue is a cancer tissue and is selected from the group consisting hepatocellular carcinoma (HCC), colorectal carcinoma (CRC), glioblastoma (GB), gastric cancer (GC), esophageal cancer, non-small cell lung cancer (NSCLC), pancreatic cancer (PC), renal cell carcinoma (RCC), benign prostate hyperplasia (BPH), prostate cancer (PCA), ovarian cancer (OC), melanoma, breast cancer (BRCA), chronic lymphocytic leukemia (CLL), Merkel cell carcinoma (MCC), small cell lung cancer (SCLC), Non-Hodgkin lymphoma (NHL), acute myeloid leukemia (AML), gallbladder cancer and cholangiocarcinoma (GBC, CCC), urinary bladder cancer (UBC), uterine cancer (UEC), and combination thereof.
14. The method of any one of claims 1–13, wherein the spectral library or database comprises peptide data evaluated from over about 1500, over about 2000, over about 2500, or over about 3000 tissue samples.
15. The method of any one of claims 1–14, wherein the spectral library or database comprises over about 100 million, over about 150 million, over about 180 million, or over about 200 million MS/MS spectra.
16. The method of any one of claims 1–15, wherein the one or more tissue samples comprise tumor tissue and corresponding healthy tissue and are obtained from an individual for testing.
17. The method of any one of claims 1–16, wherein prior to (a) obtaining one or more tumor samples from an individual, the HLA-subtypes of the individual are determined.
18. The method of any one of claims 1–17, wherein the deep learning algorithm is selected from the group of pDeep, DeepMass, or PROSIT.
19. The method of any one of claims 1–18, wherein the deep learning algorithm is pDeep.
20. The method of any one of claims 1–19, further comprising a step of utilizing the identified one or more peptides in adoptive cellular therapy.
21. The method of any one of claims 1–20, wherein the one or more antigenic peptides bind to an MHC molecule.

22. The method of any one of claims 1-21, wherein the Deep Learning algorithm is multilayer perceptron neural network (MLPNN), backpropagation, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Generative Adversarial Network (GAN), Restricted Boltzmann Machine (RBM), pDeep, DeepMass, PROSIT, Deep Belief Network (DBN), or an ensemble thereof.
23. The method of any one of claims 1-22, wherein the method further comprises
- (a) acquiring retention time data for one or more antigenic peptides;
 - (b) comparing peptide retention time data to theoretical peptide retention time data in one or more public or non-public databases;
 - (c) generating a peptide spectrum match (PSM) of the one or more antigenic peptides using the retention time data,
 - (d) producing a matched spectral library or database of antigenic peptides based on steps (a) – (c),
 - (e) using a deep learning algorithm to train at least 80% of the peptide data located in the database or spectral library and testing the balance of the peptide data located in the database or library thereby producing a peptide prediction model to generate predicted peptide spectrum; and
 - (f) using the prediction model to identify one or more antigenic peptides.
24. A method classifying test data, the test data comprising peptide spectrum data, the method comprising:
- (a) receiving, on at least one processor, test data comprising peptide spectrum data,
 - (b) evaluating, using the at least one processor, the test data using a classifier which is an electronic representation of a classification system, each said classifier trained using an electronically stored set of training data vectors, each training data vector representing an individual peptide and comprising a peptide spectrum data for the peptide, each training data vector further comprising a classification with respect to whether or not the peptide is antigenic,
 - (c) outputting, using the at least one processor, a classification of the sample from the peptide spectrum data concerning the likelihood of whether or not the peptide is antigenic based on the evaluating step.
25. A method of classifying test data, the test data comprising peptide spectra data, the method comprising:
- (a) accessing, using at least one processor, an electronically stored set of training data vectors, each training data vector representing an individual cancer patient and

comprising a peptide spectrum data for the respective cancer patient, each training data vector further comprising a classification with respect to whether or not a peptide is antigenic;

- (b) training an electronic representation of a classification system, using the electronically stored set of training data vectors;
- (c) receiving, at the at least one processor, test data comprising peptide spectrum data;
- (d) evaluating, using the at least one processor, the test data using the electronic representation of the classification system; and
- (e) outputting a classification of the test data concerning whether or not the peptide is antigenic based on the evaluating step.

26. A method of classifying test data, the test data comprising peptide spectra data, the method comprising:

- (a) accessing, using at least one processor, an electronically stored set of training data vectors, each training data vector representing an individual type of cancer and comprising a peptide spectrum data for the respective type of cancer, each training data vector further comprising a classification with respect to whether or not a peptide is antigenic;
- (b) training an electronic representation of a classification system, using the electronically stored set of training data vectors;
- (c) receiving, at the at least one processor, test data comprising peptide spectrum data;
- (d) evaluating, using the at least one processor, the test data using the electronic representation of the classification system; and
- (e) outputting a classification of the test data concerning whether or not the peptide is antigenic based on the evaluating step.

27. A method of classifying test data comprising:

- (a) analyzing at least one tissue sample by mass spectrometry (MS);
- (b) acquiring experimental mass spectra data from at least one peptide bound to a HLA-subtype in a tissue sample,
- (c) generating a peptide spectrum match (PSM) of the peptide by comparing the acquired mass spectra to peptide theoretical spectra,
- d) producing a data set of peptides based on steps (a), (b), (c), and combinations thereof,
- e) evaluating the data set of peptides using a classification system to determine whether or not the peptide is antigenic.

28. A method of classifying test data comprising:

- (a) obtaining a tumor sample from an individual,
 - (b) acquiring mass spectrometry spectra for a peptide bound to HLA-subtypes in the tumor sample,
 - (c) comparing the mass spectrometry data spectra to peptide theoretical spectra located in a database,
 - (d) generating a peptide spectrum match (PSM) of the peptide,
 - (e) producing a data set of matched peptides based on steps (a), (b), (c), (d), or a combination thereof,
 - (g) evaluating the data set of peptides using a classification system to generate a predicted peptide fragmentation spectra.
- 29.** A method of classifying test data comprising:
- (a) determining HLA-subtypes of an individual,
 - (b) obtaining at least one tumor tissue sample and corresponding healthy tissue sample from an individual,
 - (c) identifying at least one peptide bound to HLA-subtypes in the tumor sample by mass spectrometry (MS),
 - (d) generating experimental spectra from the mass spectrometry data;
 - (e) comparing the mass spectrometry experimental spectra to spectra found in a database,
 - (f) generating a peptide spectrum match (PSM) of the peptide,
 - (g) producing a spectral library of peptides,
 - (h) evaluating the spectral library of peptides using a classification system to generate a peptide mass spectrometry fragmentation prediction model, and
 - (i) using the prediction model to generate predicted spectrum to identify peptides.
- 30.** A method of classifying test data to identify an antigenic peptide comprising:
- (a) obtaining at least one tumor tissue sample and corresponding healthy tissue sample from an individual,
 - (b) identifying at least one antigenic peptides bound to HLA-subtypes in a tumor tissue sample by mass spectrometry (MS), to produce a experimental peptide fragmentation spectra of the antigenic peptide;
 - (c) comparing experimental peptide fragmentation spectra to those found in a database;
 - (d) estimation of false discovery rate (FDR);
 - (e) generation of a peptide spectrum match (PSM);
 - (f) inputting the data generated by the experimental mass spectrometry methodology into a classification system to train a peptide fragmentation prediction model;

- (g) developing predicted peptide spectrum; and
 - (h) identifying an antigenic peptide.
31. A method of classifying test data to identify an antigenic peptide for use to induce a CTL immune response in an individual comprising:
- (a) determining HLA-subtypes of an individual,
 - (b) obtaining a tumor sample and a corresponding healthy tissue sample from the individual,
 - (c) identifying at least one gene that are expressed in the tumor sample relative to the healthy tissue sample,
 - (d) identifying at least one antigenic peptide bound to the HLA-subtypes in the tumor sample by mass spectrometry (MS) to produce peptide fragmentation spectra of at least one antigenic peptide,
 - (e) evaluating the HLA peptide sequence database using a classification system to produce a library of predicted peptide fragmentation spectra corresponding to the input HLA peptide sequences,
 - (f) matching the peptide fragmentation spectra with the library of predicted peptide fragmentation spectra,
 - (g) identifying the sequences of the antigenic peptide, when the peptide fragmentation spectra matches the predicted peptide fragmentation spectra that correspond to the HLA peptide sequences,
 - (h) selecting only the antigenic peptides identified in step (g), which are encoded by the gene identified in step (c), and
 - (i) synthesizing the antigenic peptides selected in step (h),
- wherein the CTL immune response is induced in the individual.
32. A method of classifying test data, the test data comprising peptide spectra data, the method comprising:
- (a) accessing, using at least one processor, an electronically stored set of training data vectors, each training data vector representing an individual type of cancer and comprising a peptide spectrum data for the respective type of cancer, each training data vector further comprising a classification with respect to whether or not a peptide is associated with an individual type of cancer;
 - (b) training an electronic representation of a classification system, using the electronically stored set of training data vectors;
 - (c) receiving, at the at least one processor, test data comprising peptide spectrum data;

- (d) evaluating, using the at least one processor, the test data using the electronic representation of the classification system; and
 - (e) outputting a classification of the test data concerning whether or not the peptide is associated with an individual type of cancer based on the evaluating step.
33. The method of any one of claims 1-32, wherein the database is a public database, non-public database, or a combination thereof.
34. The method of any one of claims 1-33, wherein the peptides are HLA-associated peptides.
35. The method of any one of claims 1-34, wherein the peptide spectra data is experimental spectrum data, predicted spectrum data, or a combination thereof.
36. The method of claim 35, wherein the peptide spectra data is experimental spectrum data.
37. The method of any one of claims 1-36, wherein the test data further comprises retention time data.
38. The method of any one of claims 1-37, wherein the gene expression is identified using microarray analysis, RNAseq, RT-PCR, RT-qPCR, or a combination thereof.
39. The method of any one of claims 1-38, wherein the classification system is AdaBoost, Artificial Neural Network (ANN) learning algorithm, Bayesian belief networks, Bayesian classifiers, Bayesian neural networks, Boosted trees, case-based reasoning, classification trees, Convolutional Neural Networks, decisions trees, Deep Learning, elastic nets, Fully Convolutional Networks (FCN), genetic algorithms, gradient boosting trees, k-nearest neighbor classifiers, LASSO, Linear Classifiers, Naïve Bayes, neural nets, penalized logistic regression, Random Forests, ridge regression, support vector machines, or an ensemble thereof.
40. The method of any one of claims 1-39, wherein the classification system is an ensemble of classification systems.
41. The method of any one of claims 1-40, wherein the classification system is Deep Learning.
42. The method of any one of claims 1-41, wherein the Deep Learning is multilayer perceptron neural network (MLPNN), backpropagation, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Generative Adversarial Network (GAN), Restricted Boltzmann Machine (RBM), Deep Belief Network (DBN), or an ensemble thereof.
43. The method of any one of claims 1-42, wherein the Deep Learning is selected from the group consisting of pDeep, Deep neural network, deep belief network, recurrent neural network, convolutional neural network, DeepMass, PROSIT, or an ensemble thereof.
44. The method of claim 43, wherein the Deep Learning is pDeep.

45. The method of any one of claims 1-44, wherein the test data is obtained from mass spectroscopy, is predicated spectra data, or a combination thereof.
46. The method of any one of claims 1-45, wherein the test data is obtained from a peptide obtained from a physiological sample from a patient, preferably a tumor biopsy.
47. The method of any one of claims 1-46, wherein the method further comprises obtaining a tissue sample and acquiring mass spectrometry spectra for the test data.
48. The method of claim 47, wherein the mass spectrometry comprises tandem mass spectrometry (MS/MS).
49. The method of any one of claims 1-48, wherein the library or database comprises over about 70%, over about 80%, over about 85%, over about 90%, over about 95%, or 100% antigenic peptide data.
50. The method of any one of claims 1-48, wherein the library or database comprises less than about 30%, less than about 25%, less than about 20%, less than about 15%, less than about 10%, or less than about 5% tryptic peptides data.
51. The method of any one of claims 1-50, wherein the antigenic peptide is identified by the predicted spectra have an identification correlation within about 2% to about 15% relative to the actual technical variation of the experimentally determined spectra.
52. The method of any one of claims 1-51, wherein the prediction peptide performance score is greater than about 0.95.
53. The method of any one of claims 1-52, wherein the prediction peptide performance score is from about 0.92 to about 0.98 or at about 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, or 0.98.
54. The method of any one of claims 1-54, wherein the peptide spectrum match (PSM) have a false discovery rate (FDR) of less than 0.05.
55. The method of any one of claims 1-54, wherein antigenic peptides are identified with greater accuracy than tryptic peptides.
56. The method of any one of claims 1-55, wherein the one or more identified antigenic peptides exhibit a peptide performance score that is closer to the measured technical variation as compared to analyzing the same peptide with ProteomeTools.
57. The method of any one of claims 1-55, wherein the antigenic peptides are 8 to 11 amino acids in length, preferably 8 or 9 amino acids in length.
58. The method of any one of claims 1-57, wherein the antigenic peptide identified is overexpressed or presented in a specific type of cancer.
59. The method of any one of claims 1-58, wherein the cancer is selected from the group consisting of hepatocellular carcinoma (HCC), colorectal carcinoma (CRC), glioblastoma

(GB), gastric cancer (GC), esophageal cancer, non-small cell lung cancer (NSCLC), pancreatic cancer (PC), renal cell carcinoma (RCC), benign prostate hyperplasia (BPH), prostate cancer (PCA), ovarian cancer (OC), melanoma, breast cancer (BRCA), chronic lymphocytic leukemia (CLL), Merkel cell carcinoma (MCC), small cell lung cancer (SCLC), Non-Hodgkin lymphoma (NHL), acute myeloid leukemia (AML), gallbladder cancer and cholangiocarcinoma (GBC, CCC), urinary bladder cancer (UBC), uterine cancer (UEC), or a combination thereof.

60. The method of any one of claims 1-59, wherein the spectral library or database comprises peptide data evaluated from over about 1500, over about 2000, over about 2500, or over about 3000 tissue samples.
61. The method of any one of claims 1-59, wherein the spectral library or database comprises over about 100 million, over about 150 million, over about 180 million, or over about 200 million MS/MS spectra.
62. The method of any one of claims 1-61, wherein the tissue sample comprises tumor tissue and corresponding healthy tissue and are obtained from an individual for testing.
63. The method of any one of claims 1-62, wherein prior to (a) obtaining the tumor sample from an individual, the HLA-subtypes of the individual are determined.
64. The method of any one of claims 1-63, wherein the training data comprises peptide data obtained from mass spectroscopy, is predicated spectra data, or a combination thereof.
65. The method of any one of claims 1-64, wherein the training data comprises peptide data obtained from a physiological sample from a patient.
66. The method of claim 65, wherein the physiological sample is a tumor biopsy.
67. The method of claim 65, wherein the physiological sample is normal tissue.
68. The method of any one of claims 1-67, wherein the training data comprises peptide data obtained from at least about 1–30 different types of cancer.
69. The method of any one of claims 1-68, wherein the training data comprises peptide data obtained from cancer samples selected from the group consisting of hepatocellular carcinoma (HCC), colorectal carcinoma (CRC), glioblastoma (GB), gastric cancer (GC), esophageal cancer, non-small cell lung cancer (NSCLC), pancreatic cancer (PC), renal cell carcinoma (RCC), benign prostate hyperplasia (BPH), prostate cancer (PCA), ovarian cancer (OC), melanoma, breast cancer (BRCA), chronic lymphocytic leukemia (CLL), Merkel cell carcinoma (MCC), small cell lung cancer (SCLC), Non-Hodgkin lymphoma (NHL), acute myeloid leukemia (AML), gallbladder cancer and cholangiocarcinoma (GBC, CCC), urinary bladder cancer (UBC), uterine cancer (UEC), or a combination thereof.

70. The method of any one of claims 1-69, wherein the training data comprises peptide data obtained from at least about 1-50 different types of normal tissue.
71. The method of any one of claims 1-70, wherein the training data comprises peptide data obtained from normal tissues samples selected from the group consisting of adipose tissue, ovary, adrenal gland, pancreas, artery, peripheral nerve tissue, bladder, pituitary gland, blood cells, placenta, bone marrow, pleura, central nervous system tissue, preferably brain or spinal cord, prostate, breast, rectum, cartilage, salivary gland, cervix, skin, colon, small intestine, preferably duodenum, jejunum, ileum, appendix, stomach, esophagus, spleen, eye, gall bladder, testis, heart, thymus, kidney, thyroid, liver, trachea, lung, ureter, lymph node, uterus, muscle, vein, or a combination thereof.
72. The method of any one of claims 1-71, wherein the training data comprises peptide data evaluated from over about 1500, over about 2000, over about 2500, or over about 3000 tissue samples.
73. The method of any one of claims 1-71, wherein the training data comprises over about 100 million, over about 150 million, over about 180 million, or over about 200 million MS/MS spectra.
74. The method of any one of claims 1-73, wherein the training data further comprises retention time data.
75. The method of any one of claims 1-74, further comprising a step of utilizing the identified one or more peptides in adoptive cellular therapy.
76. The method of any one of claims 1-75, wherein the antigenic peptide binds to an MHC molecule.
77. The method of claim 76, wherein the MHC molecule is MHC Class I or MHC Class II.
78. The method of any one of claims 1-77, wherein the data further comprises retention time.
79. A non-transitory computer readable medium storing an executable program comprising instructions to perform the method of any one of claims 1-78.
80. A system, comprising: a server comprising at least one processor and memory comprising computer-readable instructions which when executed by the processor cause the processor to perform the steps comprising: receiving peptide spectra data from a computer terminal that is located remotely from the server; processing the peptide spectra data using a classification system.