disclosure also sets forth methods and systems for automated detection of the presence of a quantity expressed in terms of, or derived from frequency-dependent data.

[0006a] According to a first aspect, the present disclosure refers to a method of identifying mass spectrometry peaks, comprising: providing at least one trained neural network model that produces an output as a function of input parameters, wherein values for the input parameters comprise a sequence of retention time values for a glycopeptide having a ~~m/z (~~mass to charge (m/z) ratio; ~~and~~ providing ~~as values for the input parameters~~ mass spectrometry (MS) data comprising ~~mass-to-charge (~~m/z~~)~~ ratios over the sequence of retention time values from a mass spectrometry of a biologic sample as the values for the input parameters; receiving, from the at least one trained neural network model,~~wherein~~ the output ~~comprises~~ comprising a location for one or more peak start and peak stop retention time values indicating ~~the~~ a presence of one or more analytes, or a fragment thereof, in the biologic sample; determining a confidence score for the identified peak start and peak stop retention time values; and excluding the identified peak start and peak stop retention time values from analysis, in response to the confidence score being below a threshold.

[0006b] According to a second aspect, the present disclosure refers to a system comprising: a data repository of mass spectrometry (MS) data; one or more processors coupled to the data repository; machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs the method of the first aspect.

[0006c] According to an example ~~third aspect~~, the present disclosure refers to a method comprising: featurizing data from raw MS data, wherein the raw MS data comprises a plurality of analog samples, each representing mass/charge intensity, the featurizing comprising, for each of the analog samples, the steps of centering the sample within a retention time window; discretizing the sample into a sequence of points representing intensity values; standardizing the intensity values; and assigning labels to points among the sequence of points corresponding to peak start and peak stop times to produce a labeled sequence of points; wherein the labeled sequence of points are configured for training a machine learning model to predict an abundance in unseen MS data.

[0006d] According to an example ~~fourth aspect~~, the present disclosure refers to a system comprising: a data repository of raw mass spectrometry (MS) data comprising a plurality of analog samples; one or more processors coupled to the repository; machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs the method of the ~~third aspect~~above example.

[0006e] According to an example ~~fourth aspect~~, the present disclosure refers to a system comprising: a data repository of raw mass spectrometry (MS) data comprising a plurality of analog samples; one or more processors coupled to the repository; machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs the method of the ~~third aspect~~above example.

[0006f] According to an example ~~fifth aspect~~, the present disclosure refers to a method comprising: computing a distribution of probabilities over a sequence of points representing mass spectrometry (MS) data for a given m/z ratio and retention time period, the probabilities representing a likely peak start or peak stop retention time; and computing a statistic for the abundance using the distribution of probabilities, wherein the statistic is indicative of the presence of a biologic sample in the MS data.

[0006g] According to a third ~~sixth~~ aspect, the present disclosure refers to a method of training a model to identify mass spectrometry peaks, comprising: using a plurality of labeled sequential data, each of the plurality of labeled sequential data comprising a peak start and a peak stop label representing retention time values characteristic of glycopeptides or peptides, or fragments thereof; ~~and~~ training at least one machine learning model using the labeled sequential data; identifying, using ~~wherein~~ the at least one trained machine learning model, ~~is adapted for identifying~~ mass spectrometry peak start and peak stop points indicating ~~the~~ a presence of one or more glycopeptides or peptides, or a fragment thereof, ~~present~~ in a biologic sample; determining a confidence score for the mass spectrometry peak start and peak stop points; and excluding the mass spectrometry peak start and peak stop points from analysis, in response to the confidence score being below a threshold.

[0006h] According to a ~~seventh~~ fourth aspect, the present disclosure refers to a method comprising computing an abundance for a glycopeptide or peptide using a machine learning model trained by the ~~process~~ method of the ~~sixth~~ third aspect.

[0006i] According to a fifth ~~eighth~~ aspect, the present disclosure refers to a system comprising a data repository of labeled sequential data; one or more processors coupled to the data repository; machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs the method of the ~~sixth~~ third and ~~seventh~~ fourth aspects.

[0006j] According to a ~~nineth~~ sixth aspect, the present disclosure refers to a method comprising: using a plurality of labeled sequential data; and training at least one machine learning model using the labeled sequential data, identifying, using~~wherein~~ the at least one trained machine learning model, ~~is adapted for identifying~~ a time period or frequency range where a peak start and peak stop peak points of an intensity occur in a~~n~~ ~~unseen~~ biologic sample; determining a confidence score for the time period or the frequency range where the peak start and peak stop points of the intensity occur; and excluding the time period or the frequency range from analysis, in response to the confidence score being below a threshold.

[0006k] According to an example ~~tenth aspect~~, the present disclosure refers to a method comprising: providing at least one trained neural network model that produces an output as a function of input parameters, wherein values for the input parameters are a temporal or spectral sequence of intensity values representing a quantity; and providing values for each of the input parameters, the values representing an unseen sequence of intensities for a quantity; wherein the output comprises one or more peak start and peak stop times or frequencies, which are used to quantify the quantity.

[0007] In one example, automated detection is applied to a time series of intensities vs. time. The intensities derive from a detector which detects the presence or absence of a quantity. For example, the automated detection may be applied to mass spectrometry (MS) in which the intensities derive from a MS detector detecting a transition value as a function of time, *e.g.*, retention time. As such, the automated detection described herein may be applied to a wide variety of data inputs based on intensity values observed as a function of time or as a function of frequency.

[0008] In one example the automated detection is applied to a time series of mass to charge ratio intensities vs. time, which were produced from a process involving mass spectrometry (MS). In another example the automated detection is applied to a time series of intensities vs. time, which were produced from a process involving liquid or gas chromatography.

[0009] According to other examples the automated detection may be applied to any of the following intensities expressed in terms of, or derived from time or frequency data:

- x-ray diffraction in which the intensities derive from an x-ray detector, *e.g.*, charge coupled device (CCD), detecting a reflection of x-rays as a function of the x-ray scan angle, *e.g.*, 2-Θ, in which scan angles are observed as a function of time.
- Raman spectroscopy in which the intensities derive from an infrared detector detecting vibrational modes as a function of light wavelength, *e.g.*, cm^{-1}, absorption or emission.
- UV-VIS spectroscopy in which the intensities derive from a light detector detecting intensities of electronic transitions as a function of ultra-violet and visible light wavelengths or frequencies. fluorimetry spectroscopy or phosphorimetry spectroscopy in which emitted fluorescence and phosphorescence light intensity is detected as a function of incident, *i.e.*, absorbed or excitation, wavelength.
- Nuclear Magnetic Resonance (NMR) spectroscopy in which the intensities derive from a radio wave detector detecting nuclear transitions as a function of radiation frequency absorption or emission within a strong magnetic field. In NMR, the intensity v. frequency detection can be converted, *e.g.*, using Fourier transformations, into an equivalent intensity v. time. The same principles applicable to NMR are also applicable to Magnetic Resonance Imaging (MRI).

WHAT IS CLAIMED IS:

1. A method of identifying mass spectrometry peaks comprising:

providing at least one trained neural network model that produces an output as a function of input parameters, wherein values for the input parameters comprise a sequence of retention time values for a glycopeptide having a ~~m/z (~~mass to charge (m/z) ratio;~~ and~~

providing ~~as values for the input parameters~~ mass spectrometry (MS) data comprising ~~mass-to-charge (~~m/z~~)~~ ratios over the sequence of retention time values from a mass spectrometry of a biologic sample as the values for the input parameters;

receiving, from the at least one trained neural network model, ~~wherein~~ the output ~~comprises~~ comprising a location for one or more peak start and peak stop retention time values indicating ~~the~~ a presence of one or more analytes, or a fragment thereof, in the biologic sample;

determining a confidence score for the identified peak start and peak stop retention time values; and

excluding the identified peak start and peak stop retention time values from analysis, in response to the confidence score being below a threshold.

2. The method of claim 1, wherein the method is a method of identifying peak start and peak stop retention time values of the mass spectrometry peaks.

3. The method of claim 2, further comprising quantifying the presence of the one or more glycopeptides or peptides by summing or integration over retention time values between the mass spectrometry peak start and peak stop ~~peaks~~ from the biologic sample.

4. The method of Claim 3, wherein the quantifying ~~step includes~~comprises applying a weighted smoothing function over the identified peak start and peak stop retention time values.

5. The method of claim 4, wherein ~~a confidence score is calculated for the identified peak start and peak stop retention time values, wherein if the value of the confidence score is below a threshold value,~~ the identified peak start and peak stop retention time values are ~~excluded from analysis and~~ targeted for reevaluation by a human expert , in response to the confidence score being below the threshold.

6. The method of any of claims 1-5, wherein the mass spectrometry data is tandem chromatography-mass spectrometry data.

7. A system comprising

a data repository of mass spectrometry (MS) data;

one or more processors coupled to the data repository;

machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs the method of any of claims 1 – 6.

~~8. A method comprising:~~

~~featurizing data from raw MS data, wherein the raw MS data comprises a plurality of analog samples, each representing mass/charge intensity, the featurizing comprising, for each of the analog samples, the steps of~~

~~centering the sample within a retention time window;~~

~~discretizing the sample into a sequence of points representing intensity values;~~

~~standardizing the intensity values; and~~

~~assigning labels to points among the sequence of points corresponding to peak start and peak stop times to produce a labeled sequence of points;~~

~~wherein the labeled sequence of points are configured for training a machine learning model to predict an abundance in unseen MS data.~~

~~9. The method of Claim 8, wherein the centering comprises positioning an apex of a curve at a center of the retention time window.~~

~~10. The method of Claim 8, wherein the standardizing the intensity values excludes intensity values greater than a threshold amount.~~

~~11. The method according to any of Claim 10, wherein the threshold amount is about 500.~~

~~12. The method of any of claims 8–11, wherein the MS data is tandem chromatography-mass spectrometry data.~~

~~13. A system comprising~~

~~a data repository of raw mass spectrometry (MS) data comprising a plurality of analog samples;~~

~~one or more processors coupled to the repository;~~

~~machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs the method of any of Claims 8—12.~~

~~14. A method comprising~~

~~computing a distribution of probabilities over a sequence of points representing mass spectrometry (MS) data for a given m/z ratio and retention time period, the probabilities representing a likely peak start or peak stop retention time; and~~

~~computing a statistic for the abundance using the distribution of probabilities, wherein the statistic is indicative of the presence of a biologic sample in the MS data.~~

~~15. The method of Claim 14, wherein the statistic is a mean abundance computed as a weighted sum of integrals using all possible pairs of start and end points for evaluating the integral.~~

~~16. The method of claim 14 or claim 15, wherein the mass spectrometry data is tandem chromatography-mass spectrometry data.~~

~~17~~8. A method of training a model to identify mass spectrometry peaks comprising:

using a plurality of labeled sequential data, each of the plurality of labeled sequential data comprising a peak start and a peak stop label representing retention time values characteristic of glycopeptides or peptides, or fragments thereof; ~~and~~

training at least one machine learning model using the labeled sequential data;

identifying, using ~~wherein~~ the at least one trained machine learning model, ~~is adapted for identifying~~ mass spectrometry peak start and peak stop points indicating ~~the~~ a presence of one or more glycopeptides or peptides, or a fragment thereof, ~~present~~ in a biologic sample;

determining a confidence score for the mass spectrometry peak start and peak stop points; and

excluding the mass spectrometry peak start and peak stop points from analysis, in response to the confidence score being below a threshold.

~~18~~9. The method of claim ~~17~~8, further comprising quantifying the presence of the one or more glycopeptides or peptides by integration over time between the mass spectrometry peak start and peak stop points from the biologic sample.

~~19~~10. The method of claim ~~17~~8, wherein the at least one machine learning model is a recurrent neural network model.

~~20~~11. The method of claim ~~19~~10, wherein the recurrent neural network is a bi-directional LSTM recurrent neural network model with attention layer.

~~21~~12. The method of any of claims ~~17~~8–~~20~~11, wherein the method is a method of training a machine learning model to identify peak start and peak stop retention time values.

~~22~~13. The method of claim ~~21~~12, wherein the mass spectrometry data is tandem chromatography-mass spectrometry data.

~~23~~14. A method comprising computing an abundance for a glycopeptide or peptide using a machine learning model trained by the ~~process~~ method of claim ~~17~~8.

~~24~~15. A system comprising

a data repository of labeled sequential data;

one or more processors coupled to the data repository;

machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs the method of any of claims ~~17~~ 8 – ~~23~~14.

~~25~~16. The method of any of Claims 1~~, 8~~ or ~~17~~8, wherein the at least one machine learning model is a sequential model (MODEL A) or a reference-based model (MODEL B).

~~26~~17. A method comprising:

using a plurality of labeled sequential data; and

training at least one machine learning model using the labeled sequential data,

identifying, using ~~wherein~~ the at least one trained machine learning model, ~~is adapted for identifying~~ a time period or frequency range where peak start and peak stop points of an intensity occur in a~~n unseen~~ biologic sample;

determining a confidence score for the time period or the frequency range where the peak start and peak stop points of the intensity occur; and

excluding the time period or the frequency range from analysis, in response to the confidence score being below a threshold.

~~27~~18. The method of Claim ~~26~~17, wherein the peak start and peak stop points are ~~contained~~ in a noise to signal transition time period.

~~28~~19. The method of Claim ~~26~~17, wherein the machine learning model is a recurrent neural network with an attention layer.

~~29~~20. The method of Claim ~~26~~17, wherein the sample represents a physical thing such that a summation or integration over time or frequency from the peak start point to the peak stop point of the intensity quantifies the physical thing.

~~30~~21. The method of Claim ~~29~~20, wherein the sample is a sequential series of mass to charge intensities over a specified time period and the summation or integration represents ~~the~~ an abundance of the physical thing.

~~31. A method comprising:~~

~~providing at least one trained neural network model that produces an output as a function of input parameters, wherein values for the input parameters are a temporal or spectral sequence of intensity values representing a quantity; and~~

~~providing values for each of the input parameters, the values representing an unseen sequence of intensities for a quantity;~~

~~wherein the output comprises one or more peak start and peak stop times or frequencies, which are used to quantify the quantity.~~

~~32. The method of Claim 31, wherein the quantity is quantified by summing or integration over time or frequency between the peak start and peak stop times or frequencies, respectively.~~

~~33. The method of claim 31 or claim 32, wherein the intensities are obtained from a process selected from the set of mass spectrometry, MRM, liquid or gas chromatography, x-ray diffraction, Raman spectroscopy, UV-VIS spectroscopy, fluorimetry spectroscopy or phosphorimetry spectroscopy, Nuclear Magnetic Resonance (NMR) spectroscopy, or Magnetic Resonance Imaging (MRI).~~

~~34. The method of any of claims 31–33, wherein the method is performed by a system comprised of~~

~~a data repository of intensity data;~~

~~one or more processors coupled to the repository;~~

~~machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs the method.~~