

SYSTEMS AND METHODS FOR GLYCOPEPTIDE CONCENTRATION DETERMINATION, NORMALIZED ABUNDANCE DETERMINATION, AND LC/MS RUN SAMPLE PREPARATION

5 [0001] This application claims priority to U.S. Provisional Patent Application Serial No. 63/251,028, filed September 30, 2021, and also claims priority to U.S. Provisional Patent Application Serial No. 63/293,506, filed December 23, 2021, both of which applications are incorporated herein in their entirety.

FIELD

10 [0002] This description is generally directed towards systems and methods for normalizing raw abundance data generated using mass spectrometry.

BACKGROUND

[0003] Protein glycosylation and other post-translational modifications play vital roles in virtually all aspects of human physiology. Unsurprisingly, faulty or altered protein glycosylation often accompanies various phenotypes (*e.g.* disease states). The identification of
15 aberrant glycosylation provides opportunities for early detection, intervention, and treatment of affected subjects.

[0004] Protein glycosylation data can be used to train predictive models for identifying phenotypes of subjects based on the subject's protein glycosylation profile. Predictive models
20 make better predictions using large quantities of high-quality data, however, making comparisons across multiple experiments or between samples can be challenging in a variety of fields, including the field of mass spectrometry. A variety of factors can make comparisons difficult, including, but not limited to variation in several of the sample preparation and sample collection steps between experiments and samples. More specifically, efficiency of trypsin
25 digest, ionization efficiency, detector efficiency, and other contributing factors can lead to a high degree of variability, thereby, making existing systems and methods for normalizing data between samples and experiments either cost prohibitive or non-existent.

[0005] In light of the above, there is a need for improved data relativization and normalization systems and methods so that high-quality, comparative data can be used in
30 building predictive models. The present disclosure addresses this and other needs by providing systems and methods for comparing glycosylation data across samples and across multiple

experiments. More specifically, the present disclosure includes systems and methods for relativizing and normalizing protein glycosylation data to better train predictive models.

[0006] There is also a further need for new sample preparation and processing methods and systems addressing the current challenges involving data normalization in existing systems.

5 Specifically, the industry needs effective standards for normalizing abundance with samples and between samples. The systems and methods described herein address those needs and more.

SUMMARY

[0007] In various aspects, a method of determining a concentration of a target glycopeptide
10 analyte in a sample is described. In various embodiments, the method can comprise measuring an abundance for a non-glycosylated endogenous peptide (NGEP) of a sample. In some embodiments, the method can comprise, determining a corrected NGEP abundance for the NGEP using an external standard. In some embodiments, the method can comprise, measuring an internal standard abundance for an internal standard having a known internal standard
15 concentration. In some embodiments, the method can comprise, determining an NGEP concentration for the NGEP as a function of the corrected NGEP abundance, the known internal standard concentration, and the internal standard abundance. In some embodiments, the method can comprise, determining a normalized abundance of a target glycopeptide analyte of the sample as a function of a measured abundance of the target glycopeptide analyte and the
20 measured abundance of the NGEP on the same protein as the target glycoprotein analyte. In some embodiments, the method can comprise determining a target glycopeptide analyte concentration as a function of the NGEP concentration and the normalized abundance.

[0008] In some aspects, a non-transitory computer-readable medium storing computer instructions that, when executed by a computer, cause the computer to perform a method for
25 determining a concentration of a target glycopeptide analyte of the sample in a sample is described. In various embodiments, the method can comprise, measuring an abundance for a non-glycosylated endogenous peptide (NGEP) in a sample. In various embodiments, the method can comprise, determining a corrected NGEP abundance for the NGEP using an external standard. In various embodiments, the method can comprise, measuring an internal
30 standard abundance for an internal standard having a known internal standard concentration. In various embodiments, the method can comprise, determining an NGEP concentration as a function of the corrected NGEP abundance, the known internal standard concentration, and the internal standard abundance. In various embodiments, the method can comprise, determining

a normalized abundance of a target glycopeptide analyte of the sample as a function of a measured abundance of the target glycopeptide analyte and the measured abundance of the non-glycosylated endogenous peptide (NGEP) on the same protein as the target glycoprotein analyte. In various embodiments, the method can comprise, determining a target glycopeptide
5 analyte concentration as a function of the NGEP concentration and the normalized abundance.

[0009] In various aspects, a method of determining a concentration of a target glycopeptide analyte in a sample is described. In various embodiments, the method can comprise measuring an abundance for a NGEP of a sample. In various embodiments, the method can comprise determining a corrected NGEP abundance for the NGEP using an external standard. In various
10 embodiments, the method can comprise measuring an internal standard abundance for an internal standard having a known internal standard concentration. In various embodiments, the method can comprise determining a NGEP concentration as a function of the corrected NGEP abundance, the known internal standard concentration, and the internal standard abundance. In various embodiments, the method can comprise determining a site occupancy of a target
15 glycopeptide analyte of the sample as a function of a measured target abundance of the target glycopeptide analyte at a given site, and a measured total abundance of all glycopeptides quantified at the same site. In various embodiments, the method can comprise determining a target glycopeptide analyte concentration as a function of the NGEP concentration and the normalized abundance.

[0010] In various aspects, a non-transitory computer-readable medium storing computer instructions that, when executed by a computer, cause the computer to perform a method for determining a concentration of a target glycopeptide analyte of a sample is described. In various
20 embodiments, the method can comprise, measuring an abundance for a non-glycosylated endogenous peptide (NGEP) of a sample. In various embodiments, the method can comprise, determining a corrected NGEP abundance for the NGEP using an external standard. In various
25 embodiments, the method can comprise, measuring an internal standard abundance for an internal standard having a known internal standard concentration. In various embodiments, the method can comprise, determining a NGEP concentration as a function of the corrected NGEP abundance, the known internal standard concentration, and the internal standard abundance. In
30 various embodiments, the method can comprise, determining a site occupancy of a target glycopeptide analyte of the sample as a function of a measured target abundance of the target glycopeptide analyte at a given site, and a measured total abundance of all glycopeptides at the

same site. In various embodiments, the method can comprise, determining a target glycopeptide analyte concentration as a function of the NGEF concentration and the normalized abundance.

[0011] In some aspects, a method for preparing samples for a liquid chromatography/mass spectrometry (LC-MS) sample run and normalizing abundance data of a plurality of sample runs is described in accordance with various embodiments. In some embodiments, the method comprises preparing run samples for the plurality of sample runs for LC-MS analysis. In various embodiments, the step of preparing comprises generating a first set of run samples. In various embodiments, each of the first set of run samples includes an external standard. In various embodiments the step of preparing comprises generating a second set of run samples. In various embodiments, each of the second set of run samples includes at least two target glycopeptide analytes and an internal standard. In various embodiments, the method comprises combining at least two run samples of the first set of run samples to create a pooled standard. In various embodiments, the method comprises analyzing each run sample according to a run order. In various embodiments, the run order specifies a relative order of analysis for the run samples of the sample run. In various embodiments, the method comprises normalizing the raw abundance data of the plurality of sample runs using the pooled standard of each sample run as a reference.

[0012] In some aspects, a sample processing system for carrying out a process for preparing samples for a liquid chromatography/mass spectrometry (LC-MS) sample run and normalizing abundance data of a plurality of sample runs is described in accordance with various embodiments. In various embodiments, the sample processing system comprises a sample preparation system. In various embodiments, the sample preparation system includes a fluidic instrument for performing a process. In various embodiments, the process can comprise generating a first set of run samples. In various embodiments, each of the first set includes an external standard. In various embodiments, the process can comprise generating a second set of run samples. In various embodiments, each of the second set includes at least two target glycopeptide analytes and an internal standard. In various embodiments, the process can comprise combining at least two run samples of the first set to create a pooled standard. In various embodiments, the sample processing system can comprise a sample analysis system. In various embodiments, the sample analysis system can include a LC-MS instrument for analyzing glycopeptide analytes. In various embodiments, the sample analysis system can analyze each run sample according to a run order. In various embodiments, the run order can be stored on a data store in electronical communication with the LC-MS instrument and

specifies a relative order of analysis for the run samples of the sample run. In various embodiments, the sample processing system comprises a data analysis system, including a normalization module of a peptide structure analyzer, for normalizing raw abundance data of the plurality of sample runs using the pooled standard of each sample run as a common reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] The present disclosure is described in conjunction with the appended figures:

[0014] Figure 1 is a schematic diagram of an exemplary workflow 100 for the detection of peptide structures associated with a disease state for use in diagnosis and/or treatment in accordance with one or more embodiments.

[0015] Figure 2A is a schematic diagram of a preparation workflow in accordance with one or more embodiments.

[0016] Figure 2B is a schematic diagram of data acquisition in accordance with one or more embodiments.

[0017] Figure 3 is a block diagram of an analysis system in accordance with one or more embodiments.

[0018] Figure 4 is a block diagram of a computer system in accordance with various embodiments.

[0019] Figure 5 is flowchart of a process for determining a concentration of a target glycopeptide analyte in a sample in accordance with various embodiments.

[0020] Figure 6 is a flowchart of a process for determining a concentration of a target glycopeptide analyte in a sample in accordance with various embodiments.

[0021] Figure 7A illustrates experimental data showing a normalization result that does not employ combined external standards combined with internal standard and relativistic approaches. For each marker, the top horizontal line is SS-JAR1.P2.R3, the middle horizontal line is SS-JAR1.P1.R1, and the bottom horizontal line is HVI-CRC.P4.R6.

[0022] Figure 7B illustrates experimental data showing a normalization resulting by adding an additional step of abundance correction *via* external standard to the combined approaches in 7A. For each marker, the top horizontal line is SS-JAR1.P2.R3, the middle horizontal line is SS-JAR1.P1.R1, and the bottom horizontal line is HVI-CRC.Ab.Test.

[0023] Figure 8A illustrates hypothetical abundance values for three experiments prior to applying normalization methods described below and herein according to various embodiments.

[0024] Figure 8B illustrates hypothetical abundance values for three experiments after applying normalization methods described below and herein according to various embodiments.

5 [0025] Figure 9 is a flowchart of a process for determining a concentration of a target glycopeptide analyte in a sample is described in accordance with various embodiments.

[0026] Figure 10A illustrates experimental data where process Normalize Data Using External Standard occurs prior to processes Normalize Data Using Internal Standard and Site Occupancy Abundance and process Concentration Determination occurs last.

10 [0027] Figure 10B illustrates the same experimental data from Figure 10A where process Normalize Data Using External Standard does not occur.

[0028] Figure 11A illustrates experimental data where process Normalize Data Using External Standard occurs prior to processes Normalize Data Using Internal Standard and Relative Abundance and process Concentration Determination occurs last.

15 [0029] Figure 11B illustrates the same experimental data from Figure 11A where process Normalize Data Using External Standard does not occur.

[0030] Figure 12A and Figure 12B illustrate confusion matrices resulting in applying a trained ovarian cancer model to an independent test set and demonstrate an increase in predictive accuracy using the systems and methods described herein and in accordance with various embodiments.

20 [0031] Figure 13 illustrates a flowchart of an exemplary method for preparing samples for a LC-MS sample run and normalizing abundance data of a plurality of sample runs.

[0032] Figure 14 illustrates an exemplary sample processing system for processing patient sample into normalized abundance.

25 [0033] Figure 15 illustrates experimental results for data including coefficient of variation for pooled replicates.

[0034] Figure 16 illustrates an experimental run order for a set of run samples.

[0035] Figure 17 illustrates a layout for a 96-well plate including run sample partition locations.

DETAILED DESCRIPTION

30 I. Overview

[0036] The embodiments described herein recognize that glycoproteomics is an emerging field that can be used in the overall diagnosis and/or treatment of subjects with various types of diseases. Glycoproteomics aims to determine the positions, identities, and quantities of

glycans and glycosylated proteins in a given sample (*e.g.*, blood sample, cell, tissue, *etc.*). Protein glycosylation is one of the most common and most complex forms of post-translational protein modification, and can affect protein structure, conformation, and function. For example, glycoproteins may play crucial roles in important biological processes such as cell signaling, host–pathogen interactions, and immune response and disease. Glycoproteins may therefore be important to diagnosing different types of diseases or disease states.

[0037] In using this emerging technology to its full potential, supplying large quantities of high-quality abundance data to a predictive model can improve the predictive accuracy of the model. However, raw glycopeptide abundance data may be difficult to compare across different experiments or even samples within the same experiment. The skilled artisan can appreciate that improved relativizing and normalizing data systems and methods can be a way to manage such variation. Additionally, sample preparation and processing systems and methods are necessary to generate optimal data for undergoing data analysis.

[0038] Accordingly, the embodiments described herein provide various methods and systems sample preparation and processing, including, various normalization systems and methods.

[0039] The description below provides exemplary implementations of the methods and systems described herein for the data processing. Descriptions and examples of various terms, as used herein, are provided in **Section II** below.

II. Exemplary Descriptions of Terms

[0040] As used herein, “abundance,” may refer to a quantitative value generated using mass spectrometry. In various embodiments, the quantitative value may relate to the amount of a particular peptide structure. In some embodiments, the quantitative value may comprise an amount of an ion produced using mass spectrometry. In some embodiments, the quantitative value may be expressed as an *m/z* value. In other embodiments, the quantitative value may be expressed in atomic mass units.

[0041] The term “alkylation,” as used herein, generally refers to the transfer of an alkyl group from one molecule to another. In various embodiments, alkylation is used to react with reduced cysteines to prevent the re-formation of disulfide bonds after reduction has been performed.

[0042] The term “amino acid,” as used herein, generally refers to any organic compound that includes an amino group (*e.g.* -NH₂), a carboxyl group (-COOH), and a side chain group

(R) which varies based on a specific amino acid. Amino acids can be linked using peptide bonds.

[0043] As used herein, an “artificial neural network” or “neural network” (NN) may refer to mathematical algorithms or computational models that mimic an interconnected group of artificial nodes or neurons that processes information based on a connectionistic approach to computation. Neural networks, which may also be referred to as neural nets, can employ one or more layers of nonlinear units to predict an output for a received input. Some neural networks include one or more hidden layers in addition to an output layer. The output of each hidden layer is used as input to the next layer in the network, *i.e.*, the next hidden layer or the output layer. Each layer of the network generates an output from a received input in accordance with current values of a respective set of parameters. In the various embodiments, a reference to a “neural network” may be a reference to one or more neural networks.

[0044] A neural network may process information in two ways: when it is being trained, it is in training mode and when it puts what it has learned into practice it is in inference (or prediction) mode. Neural networks learn through a feedback process (*e.g.*, backpropagation) which allows the network to adjust the weight factors (modifying its behavior) of the individual nodes in the intermediate hidden layers so that the output matches the outputs of the training data. In other words, a neural network learns by being fed training data (learning examples) and eventually learns how to reach the correct output, even when it is presented with a new range or set of inputs. A neural network may include, for example, without limitation, at least one of a Feedforward Neural Network (FNN), a Recurrent Neural Network (RNN), a Modular Neural Network (MNN), a Convolutional Neural Network (CNN), a Residual Neural Network (ResNet), an Ordinary Differential Equations Neural Networks (neural-ODE), or another type of neural network.

[0045] As used herein, the phrase “at least one of,” when used with a list of items, means different combinations of one or more of the listed items may be used and only one of the items in the list may be needed. The item may be a particular object, thing, step, operation, process, or category. In other words, “at least one of” means any combination of items or number of items may be used from the list, but not all of the items in the list may be required. For example, without limitation, “at least one of item A, item B, or item C” means item A; item A and item B; item B; item A, item B, and item C; item B and item C; or item A and C. In some cases, “at least one of item A, item B, or item C” means, but is not limited to, two of item A, one of item B, and ten of item C; four of item B and seven of item C; or some other suitable combination.

[0046] The term “biomarker,” as used herein, generally refers to any measurable substance taken as a sample from a subject whose presence is indicative of some phenomenon. Non-limiting examples of such phenomenon can include a healthy state, a disease state, a condition, or exposure to a compound or environmental condition. In various embodiments described
5 herein, biomarkers may be used for diagnostic purposes. Biomarkers can be used to diagnose a state of a subject such as, for non-limiting examples, a healthy state or a disease state.

[0047] The term “denaturation,” as used herein, generally refers to any molecule that loses quaternary structure, tertiary structure, and secondary structure which is present in their native state. Non-limiting examples include proteins or nucleic acids being exposed to an external
10 compound or environmental condition such as acid, base, temperature, pressure, radiation, *etc.*

[0048] The term “denatured protein,” as used herein, generally refers to a protein that loses quaternary structure, tertiary structure, and secondary structure which is present in their native state.

[0049] The terms “digestion” or “enzymatic digestion,” as used herein, generally refer to
15 breaking apart a polymer (*e.g.* cutting a polypeptide at a cut site). Proteins may be digested in preparation for mass spectrometry using trypsin digestion protocols. Proteins may be digested using other proteases in preparation for mass spectrometry if access is limited to cleavage sites.

[0050] The term “disease state” as used herein, generally refers to a condition that affects the structure or function of an organism. Non-limiting examples of causes of disease states may
20 include pathogens, immune system dysfunctions, cell damage caused by aging, cell damage caused by other factors (*e.g.* trauma and cancer). Disease states can include any state of a disease whether symptomatic or asymptomatic. Disease states can include disease stages of a disease progression. Disease states can cause minor, moderate, or severe disruptions in structure or function of an organism (*e.g.* a subject).

[0051] As used herein, “external standard,” may refer to a standard used to normalize data
25 from one sample run to another sample run. In various embodiments, an external standard can include a NGEP. A run sample can include an external standard. An external standard can include the same or a similar structure to the target glycopeptide analyte. An external standard source may include multiple external standards. External standards can be run independently
30 of patient samples. External standard sources can be added to a run sample for sample processing and data collection. For example, a processing step may include enzymatic digestion of one or more glycoproteins using enzymatic digestion to produce one or more external standards or external standard analytes. A sample run can include a plurality of run

samples including an external standard. Abundance data collected from external standards can be used to normalize raw abundance originating from target glycopeptide analytes of the run samples. Portions of a plurality of run samples including the external standard can be pooled to generate a pooled standard. Serum can be a source for external standards and can be available
5 through commercial suppliers (*e.g.*, Sigma-Aldrich™). Pooled standard can be used to compare run samples between sample runs.

[0052] The terms “glycan,” as used herein, generally refers to a glycoconjugate such as a carbohydrate that is covalently linked to another biological molecule. Carbohydrates can include monosaccharides, disaccharides, oligosaccharides, or polysaccharides. The biological
10 molecule can include a residue of an amino acid of a protein or polypeptide. Glycans can be covalently attached to a side chain (*i.e.* R group) of an amino acid residue of a glycopeptide.

[0053] The terms “glycopeptide” or “glycopolypeptide” as used herein, generally refers to a peptide or polypeptide comprising at least one glycan located at one or more residues of the peptide or polypeptide.

[0054] The term “glycoprotein,” as used herein, generally refers to a protein having at least one glycan residue bonded thereto. In some examples, a glycoprotein is a protein with at least one oligosaccharide chain covalently bonded thereto. Examples of glycoproteins, include but are not limited to apolipoprotein C-III (APOC3), alpha-1-antichymotrypsin (AACT), afamin (AFAM), alpha-1-acid glycoprotein 1 & 2 (AGP12), apolipoprotein B-100 (APOB),
20 apolipoprotein D (APOD), complement C1s subcomponent (C1S), calpain-3 (CAN3), clusterin (CLUS), complement component C8A chain (CO8A), alpha-2-HS-glycoprotein (FETUA), haptoglobin (HPT), immunoglobulin heavy constant gamma 1 (IgG1), immunoglobulin J chain (IgJ), plasma kallikrein (KLKB1), serum paraoxonase/arylesterase 1 (PON1), prothrombin (THRB), serotransferrin (TRFE), protein unc-13 homologA (UN13A), and zinc-alpha-2-glycoprotein (ZA2G). A glycopeptide, as used herein, refers to a fragment of a glycoprotein, unless specified otherwise to the contrary.

[0055] As used herein, an “internal standard,” may refer to a molecule that can be contained (*e.g.* added-to or spiked-into) in the same sample as a target glycopeptide analyte undergoing mass spectrometry analysis. Internal standards can be spiked into all patient samples. Internal
30 standards can also be spiked into run samples including external standards. Many internal standards can include peptides or glycopeptides. Some peptides can include glycosylation sites. For various applications, peptides used as an internal standard can be non-glycosylated at one or more glycosylation sites. Internal standards can be used for calibration purposes. Internal

standards can be used for normalization and quantification. The systems and methods herein can use non-glycosylated peptides and/or non-glycosylated endogenous peptides. When collecting data that may need to be normalized and quantified, each target glycosylation site can have an internal standard for comparison. In many cases, a glycoproteins being analyzed
5 may include a corresponding aglycosylated version for comparison. In place of an internal standard having the same structure as a target glycopeptide analyte, an surrogate internal standard can be used for the target glycopeptide analyte, wherein the internal standard includes a similar m/z ratio and or retention time to the target glycopeptide analyte. In various embodiments, internal standards are spiked into all run samples comprising glycopeptide
10 analytes undergoing mass spectrometry interrogation.

[0056] The term “linking site” or “glycosylation site” as used herein generally refers to the location where a sugar molecule of a glycan or glycan structure is directly bound (*e.g.* covalently bound) to an amino acid of a peptide, a polypeptide, or a protein. For example, the linking site may be an amino acid residue and a glycan structure may be linked *via* an atom of
15 the amino acid residue. Non-limiting examples of types of glycosylation can include N-linked glycosylation, O-linked glycosylation, C-linked glycosylation, S-linked glycosylation, and glycation.

[0057] The term “liquid chromatography,” as used herein, generally refers to a technique used to separate a sample into parts. Liquid chromatography can be used to separate, identify,
20 and quantify components.

[0058] As used herein, “machine learning” may be the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world. Machine learning uses algorithms that can learn from data without relying on rules-based programming. A machine learning algorithm may include a parametric model, a nonparametric
25 model, a deep learning model, a neural network, a linear discriminant analysis model, a quadratic discriminant analysis model, a support vector machine, a random forest algorithm, a nearest neighbor algorithm, a combined discriminant analysis model, a k-means clustering algorithm, a supervised model, an unsupervised model, logistic regression model, a multivariable regression model, a penalized multivariable regression model, or another type of
30 model.

[0059] The term “mass spectrometry,” as used herein, generally refers to an analytical technique used to identify molecules. In various embodiments described herein, mass spectrometry can be involved in characterization and sequencing of proteins.

[0060] As used herein, a “model” may include one or more algorithms, one or more mathematical techniques, one or more machine learning algorithms, or a combination thereof.

[0061] The term “m/z” or “mass-to-charge ratio” as used herein, generally refers to an output value from a mass spectrometry instrument. In various embodiments, m/z can represent a relationship between the mass of a given ion and the number of elementary charges that it carries. The “m” in m/z stands for mass and the “z” standards for charge number of ions. In some embodiments, m/z can be displayed on an x-axis of a mass spectrum.

[0062] As used herein, a “non-glycosylated endogenous peptide” (“NGEP”), may refer to a peptide structure including glycosylation sites at a location of one or more amino acids. In various embodiments, one or more glycosylation sites can be non-glycosylated (*e.g.*, do not include a glycan molecule) for a given NGEP. A NGEP and a target glycopeptide analyte can originate from the same subject. A NGEP and a target glycopeptide analyte can originate from the same protein. A NGEP and a target glycopeptide analyte can originate from the different proteins. Peptide structures, including NGEPs, can be labeled with an isotope in preparation for mass spectrometry analysis. NGEPs can be used as internal standards for the systems and methods described herein.

[0063] The term “ones” means more than one.

[0064] The terms “peptides” or “polypeptide” as used herein, can be used interchangeably, and generally refer to amino acids linked by peptide bonds. Peptides can include amino acid chains between 10 and 50 residues. Peptides can include amino acid chains shorter than 10 residues, including, oligopeptides, dipeptides, tripeptides, and tetrapeptides. Peptides can include chains longer than 50 residues and may be referred to as “proteins.” Proteins may be digested in preparation for mass spectrometry using trypsin digestion protocols. Protocols include digesting protein using proteases in preparation for mass spectrometry. Proteases can be used if access to cleavage sites is limited.

[0065] The term “peptide structure,” as used herein, generally refers to peptides or a portion thereof or glycopeptides or a portion thereof. In various embodiments described herein, a peptide structure can include any molecule comprising at least two amino acids in sequence.

[0066] As used herein, the term “plurality” may be 2, 3, 4, 5, 6, 7, 8, 9, 10, or more.

[0067] As used herein, a “peptide structure data,” may refer to any data of or relating to a peptide from a resulting mass spectrometry run. A peptide data set can comprise data obtained from a run sample including a sample and/or one or more standards. Peptide structure data can include raw abundance and/or abundance data for external standards, internal standards, and/or

target glycopeptide analytes. In some cases, target glycopeptide analytes can originate from a patient sample.

[0068] As used herein, a “pooled standard,” may refer to a standard created by combining portions of plurality of external standard digests from plurality of run samples. Abundance data from pooled standards originating from different samples runs can be used to normalize abundance data between sample runs, thereby, allowing comparison of glycopeptide analyte abundance data collected at different time points and/or from different experiments.

[0069] In many sample runs, one or more individual digestion replicates can be added to plurality of run samples and the abundance data generated can be used to normalize glycopeptide analyte data across a sample run. Run sample data originating from individual digestion replicates can be used to one or more runs including the pooled serum digest (*e.g.*, plurality of individual digestion replicates combined). Many sample runs can include one or more pooled serum digests analyzed at the end of a run order. The abundances from the pooled digests can be compared to the abundances of the one or more individual digestion replicates to normalize raw abundance data for a sample run.

[0070] The term “reduction,” as used herein, generally refers to the gain of an electron by a substance. In various embodiments described herein, a sugar can directly bind to a protein, thereby, reducing the amino acid to which it binds. Such reducing reactions can occur in glycosylation. In various embodiments, reduction may be used to break disulfide bonds between two cysteines.

[0071] As used herein, “relative abundance,” may refer to a relationship between plurality of abundances. In some cases, relative abundance can include a comparison of a proportion of one peptide structure to another peptide structure. In some cases, relative abundance can include comparing a proportion of a peptide structure to a total number of peptide structures without having actual values. In some embodiments, the comparison may comprise comparing one peptide glycoform (*e.g.* two identical peptides differing by one or more glycans) to a set of peptide glycoforms. In some embodiments, the comparison may comprise comparing a number of ions having a particular m/z ratio by a total number of ions detected. In various embodiments, a relative abundance can be expressed as a ratio. In other embodiments, a relative abundance can be expressed as a percentage. Relative abundance can be presented on a y-axis of a mass spectrum plot.

[0072] The term “sequence,” as used herein, generally refers to a biological sequence including one-dimensional monomers that can be assembled to generate a polymer. Non-

limiting examples of sequences include nucleotide sequences (*e.g.* ssDNA, dsDNA, and RNA), amino acid sequences (*e.g.* proteins, peptides, and polypeptides), and carbohydrates (*e.g.* compounds including $C_m (H_2O)_n$).

[0073] The terms “sample,” “biological sample,” “biological specimen,” or “biospecimen” as used herein, generally refers to a specimen taken by sampling so as to be representative of the source of the specimen, typically, from a subject. The sample can be representative of an organism as a whole, specific tissue, cell type, or category or sub-category of interest. The sample may include blood. The blood of the sample can include whole blood. The blood of the sample can include plasma. The sample can include a macromolecule. The sample can include a small molecule. The sample can include a virus. The sample can include a cell or derivative of a cell. The sample can include an organelle. The sample can include a cell nucleus. The sample can include a rare cell from a population of cells. The sample can include any type of cell, including without limitation prokaryotic cells, eukaryotic cells, bacterial, fungal, plant, mammalian, or other animal cell type, mycoplasmas, normal tissue cells, tumor cells, or any other cell type, whether derived from single cell or multicellular organisms. The sample can include a constituent of a cell. The sample can include nucleotides (*e.g.* ssDNA, dsDNA, RNA), organelles, amino acids, peptides, proteins, carbohydrates, glycoproteins, or any combination thereof. The sample can include a matrix (*e.g.*, a gel or polymer matrix) comprising a cell or one or more constituents from a cell (*e.g.*, cell bead), such as DNA, RNA, organelles, proteins, or any combination thereof, from the cell. The sample may be obtained from a tissue of a subject. The sample can include a cell. Such cells may or may not include a cell wall or cell membrane. The sample can include one or more constituents of a cell but may not include other constituents of the cell. An example of such constituents may include a nucleus or an organelle. The biological sample may include a live cell. The live cell can be capable of being cultured.

[0074] The sample may be derived from another sample. The sample may include a tissue sample, such as a biopsy, core biopsy, needle aspirate, or fine needle aspirate. The sample may include a fluid sample, such as a blood sample, urine sample, or saliva sample. The sample may include a skin sample. The sample may include a cheek swab. The sample may include a cell-free or cell free sample. A cell-free sample may include extracellular polynucleotides. The sample may originate from blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, or tears. The sample may originate from red blood cells or white blood cells. The sample may originate from feces, spinal fluid, CNS fluid, gastric fluid, amniotic fluid, cyst fluid, peritoneal fluid, marrow, bile, other body fluids, tissue obtained from a biopsy, skin, or hair.

[0075] As used herein, “sample run,” refers to a grouping or set of one or more run samples scheduled to undergo mass spectrometry analysis. For example, at the beginning of a sample run, run samples including material for calibration or instrument maintenance (*e.g.*, bovine serum albumin (BSA), standards, or blanks) can be processed or/and analyzed by a mass spectrometry instrument for the analysis. Run samples can include patient samples combined with internal standards. Run samples can include external standards. Run samples can be stored in vials or partitions of plates (*e.g.*, a partition of a 96-well plate). In some cases, an automated system including a pipette robot can be programmed to process run samples of a sample run according to a run order.

[0076] As used herein, the term “set of” means one or more. For example, a set of items includes one or more items.

[0077] The term “subject,” as used herein, generally refers to an animal, such as a mammal (*e.g.*, human) or avian (*e.g.*, bird), or other organism, such as a plant. For example, the subject can include a vertebrate, a mammal, a rodent (*e.g.*, a mouse), a primate, a simian or a human.

Animals may include, but are not limited to, farm animals, sport animals, and pets. A subject can include a healthy or asymptomatic individual, an individual that has or is suspected of having a disease (*e.g.*, cancer) or a pre-disposition to the disease, and/or an individual that is in need of therapy or suspected of needing therapy. A subject can be a patient. A subject can include a microorganism or microbe (*e.g.*, bacteria, fungi, archaea, viruses). A sample, or constituent thereof, from a subject may be undergo mass spectrometry analysis, thereby, generating peptide structure data for one or more glycopeptides of the sample.

[0078] As used herein, a “target glycopeptide analyte,” may refer to a peptide structure (*e.g.*, glycosylated or aglycosylated/non-glycosylated) or a fraction of a peptide structure. Target analytes can include sub-structure associated with one or more residues of peptide sequence of the peptide structure (*e.g.* a glycan associated with a glycosylation site). Target glycopeptide analytes can include a product of one or more of the above listed structures and/or sub-structures. Target glycopeptide analytes can include or be associated with one or more detection molecule (*e.g.* signal molecule, label, or tag). Target glycopeptide analytes can be prepared using the sample preparation methods and systems described herein. Target glycopeptide analytes can be analyzed with a mass spectrometry system generating raw abundance data for the glycopeptide analyte which can later be normalized using the methods and systems described herein.

[0079] The term “training data,” as used herein generally refers to data that can be input into models, statistical models, algorithms and any system or process able to use existing data to make predictions. Training data can be normalized using the systems and methods described here.

5 [0080] As used herein, a “transition,” may refer to or identify a peptide structure. In some embodiments, a transition can refer to the specific pair of m/z values associated with a precursor ion and a product or fragment ion.

III. Overview of Exemplary Workflow

[0081] Figure 1 is a schematic diagram of an exemplary workflow 100 for the detection of peptide structures associated with a disease state for use in diagnosis and/or treatment in accordance with one or more embodiments. Workflow 100 may include various operations including, for example, sample collection 102, sample intake 104, sample preparation and processing 106, data analysis 108, and output generation 110.

[0082] Sample collection 102 may include, for example, obtaining a biological sample 112 of one or more subjects, such as subject 114. Biological sample 112 may take the form of a specimen obtained *via* one or more sampling methods. Biological sample 112 may be representative of subject 114 as a whole or of a specific tissue, cell type, or other category or sub-category of interest. Biological sample 112 may be obtained in any of a number of different ways. In various embodiments, biological sample 112 includes whole blood sample 116 obtained *via* a blood draw. In other embodiments, biological sample 112 includes set of aliquoted samples 118 that includes, for example, a serum sample, a plasma sample, a blood cell (*e.g.*, white blood cell (WBC), red blood cell (RBC) sample, another type of sample, or a combination thereof. Biological samples 112 may include nucleotides (*e.g.* ssDNA, dsDNA, RNA), organelles, amino acids, peptides, proteins, carbohydrates, glycoproteins, or any combination thereof.

[0083] In various embodiments, a single run can analyze a sample comprising an external standard (*e.g.* NGEF), an internal standard, and a glycopeptide analyte. As such, abundance values (*e.g.* abundance or raw abundance) for external standard, internal standard, and target glycopeptide analyte can be determined by mass spectrometry in the same run.

30 [0084] In various embodiments, serum samples may be analyzed prior to analyzing experimental samples. In various embodiments, serum samples can be run independently between experimental samples. In some embodiments, serum samples can be analyzed after

every 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, or more experiments. In various embodiments, serum sample data can be used in some or all of the normalization systems and methods described herein. In additional embodiments, blank samples may be processed to prevent column fouling.

5 **[0085]** Sample intake 104 may include one or more various operations such as, for example, aliquoting, registering, processing, storing, thawing, and/or other types of operations. In one or more embodiments, when biological sample 112 includes whole blood sample 116, sample intake 104 includes aliquoting whole blood sample 116 to form a set of aliquoted samples that can then be sub-aliquoted to form set of samples 120.

10 **[0086]** Sample preparation and processing 106 may include, for example, one or more operations to form set of peptide structures 122. In various embodiments, set of peptide structures 122 may include various fragments of unfolded proteins that have undergone digestion and may be ready for analysis.

15 **[0087]** Further, sample preparation and processing 106 may include, for example, data acquisition 124 based on set of peptide structures 122. For example, data acquisition 124 may include use of, for example, but is not limited to, a liquid chromatography/mass spectrometry (LC/MS) system.

20 **[0088]** Data analysis 108 may include, for example, peptide structure analysis 126. In some embodiments, data analysis 108 also includes output generation 110. In other embodiments, output generation 110 may be considered a separate operation from data analysis 108. Output generation 110 may include, for example, generating final output 128 based on the results of peptide structure analysis 126. Final output 128 may be used for determining research, diagnosis, and/or treatment.

25 **[0089]** In various embodiments, final output 128 is comprised of one or more outputs. Final output 128 may take various forms. For example, final output 128 may be a report that includes, for example, a diagnosis output, a treatment output (*e.g.*, a treatment design output, a treatment plan output, or combination thereof), analyzed data (*e.g.* relativized and normalized) or combination thereof. In some embodiments, report can comprise a target glycopeptide analyte concentration as a function of the NGEP concentration value and the normalized abundance value. In some embodiments, final output 128 may be an alert (*e.g.*, a visual alert, an audible alert, *etc.*), a notification (*e.g.*, a visual notification, an audible notification, an email notification, *etc.*), an email output, or a combination thereof. In some embodiments, final output 128 may be sent to remote system 130 for processing. Remote system 130 may include,

30

for example, a computer system, a server, a processor, a cloud computing platform, cloud storage, a laptop, a tablet, a smartphone, some other type of mobile computing device, or a combination thereof.

5 [0090] In other embodiments, workflow 100 may optionally exclude one or more of the operations described herein and/or may optionally include one or more other steps or operations other than those described herein (*e.g.*, in addition to and/or instead of those described herein). Accordingly, workflow 100 may be implemented in any of a number of different ways for use in the research, diagnosis, and/or treatment of a disease state.

IV. Detection and Quantification of Peptide Structures

10 [0091] Figures 2A and 2B are schematic diagrams of a workflow for sample preparation and processing 106 in accordance with one or more embodiments. Figures 2A and 2B are described with continuing reference to Figure 1. Sample preparation and processing 106 may include, for example, preparation workflow 200 shown in Figure 2A and data acquisition 124 shown in Figure 2B.

15

IV.A. Sample Preparation and Processing

[0092] Figure 2A is a schematic diagram of preparation workflow 200 in accordance with one or more embodiments. Preparation workflow 200 may be used to prepare a sample, such as a sample of set of samples 120 in Figure 1, for analysis *via* data acquisition 124. For
20 example, this analysis may be performed *via* mass spectrometry (*e.g.* LC-MS). In various embodiments, preparation workflow 200 may include denaturation and reduction 202, alkylation 204, and digestion 206. All areas of the preparation workflow can cause inconsistency between different samples and different experiments, necessitating the improved normalization systems and methods described herein and throughout.

25 [0093] In general, polymers, such as proteins, in their native form, can fold to include secondary, tertiary, and/or other higher order structures. Such higher order structures may functionalize proteins to complete tasks (*e.g.* enable enzymatic activity) in a subject. Further, such higher order structures of polymers may be maintained *via* various interactions between side chains of amino acids within the polymers. Such interactions can include ionic bonding,
30 hydrophobic interactions, hydrogen bonding, and disulfide linkages between cysteine residues. However, when using analytic systems and methods, including mass spectrometry, unfolding such polymers (*e.g.* peptide/protein molecules) may be desired to obtain sequence information.

In some embodiments, unfolding a polymer may include denaturing the polymer, which may include, for example, linearizing the polymer.

[0094] In one or more embodiments, denaturation and reduction 202 can be used to disrupt higher order structures (*e.g.*, secondary, tertiary, quaternary, *etc.*) of one or more proteins (*e.g.*, polypeptides and peptides) in a sample (*e.g.*, one of set of samples 120 in Figure 1). Denaturation and reduction 202 includes, for example, a denaturation procedure and a reduction procedure. In some embodiments, the denaturation procedure may be performed using, for example, thermal denaturation, where heat is used as a denaturing agent. The thermal denaturation can disrupt ionic bonding, hydrophobic interactions, and/or hydrogen bonding.

10 [0095] In one or more embodiments, the denaturation procedure may include using one or more denaturing agents in combination with heat. These one or more denaturing agents may include, for example, but are not limited to, any number of chaotropic salts (*e.g.*, urea, guanidine), surfactants (*e.g.*, sodium dodecyl sulfate (SDS), beta octyl glucoside, Triton X-100), or combination thereof. In some cases, such denaturing agents may be used in combination with heat when sample preparation workflow further includes a cleanup procedure.

[0096] The resulting one or more denatured (*e.g.*, unfolded, linearized) proteins may then undergo further processing in preparation of analysis. For example, a reduction procedure may be performed in which one or more reducing agents are applied. A reducing agent may take the form of, for example, without limitation, dithiothreitol (DTT), tris(2-carboxyethyl)phosphine (TCEP), or some other reducing agent. The reducing agent may reduce (*e.g.*, cleave) the disulfide linkages between cysteine residues of the one or more denatured proteins to form one or more reduced proteins.

25 [0097] In various embodiments, the one or more reduced proteins resulting from denaturation and reduction 202 may undergo a process to prevent the reformation of disulfide linkages between, for example, the cysteine residues of the one or more reduced proteins. This process may be implemented using alkylation 204 to form one or more alkylated proteins. For example, alkylation 204 may be used add an acetamide group to a sulfur on each cysteine residue to prevent disulfide linkages from reforming. In various embodiments, an acetamide group can be added by reacting one or more alkylating agents with a reduced protein. The one or more alkylating agents may include, for example, one or more acetamide salts. An alkylating agent may take the form of, for example, iodoacetamide (IAA), 2-chloroacetamide, some other type of acetamide salt, or some other type of alkylating agent.

[0098] In some embodiments, alkylation 204 may include a quenching procedure. The quenching procedure may be performed using one or more reducing agents (*e.g.*, one or more of the reducing agents described above).

[0099] In various embodiments, the one or more alkylated formed *via* alkylation 204 can then undergo digestion 206 in preparation for analysis (*e.g.*, mass spectrometry analysis). Digestion 206 of a protein may include cleaving the protein at or around one or more cleavage sites (*e.g.*, site 205 which may be one or more amino acid residues). For example, without limitation, an alkylated protein may be cleaved at the carboxyl side of the lysine or arginine residues. This type of cleavage may break the protein into various segments, which include one or more peptide structures (*e.g.*, glycosylated or aglycosylated).

[0100] In various embodiments, digestion 206 is performed using one or more proteolysis catalysts. For example, an enzyme can be used in digestion 206. In some embodiments, the enzyme takes the form of trypsin. In other embodiments, one or more other types of enzymes (*e.g.*, proteases) may be used in addition to or in place of trypsin. These one or more other enzymes include, but are not limited to, LysC, LysN, AspN, GluC, and ArgC. In some embodiments, digestion 206 may be performed using tosyl phenylalanyl chloromethyl ketone (TPCK)-treated trypsin, one or more engineered forms of trypsin, one or more other formulations of trypsin, or a combination thereof. In some embodiments, digestion 206 may be performed in multiple steps, with each involving the use of one or more digestion agents. For example, a secondary digestion, tertiary digestion, *etc.* may be performed. In one or more embodiments, trypsin is used to digest serum samples. In one or more embodiments, trypsin/LysC cocktails are used to digest plasma samples.

[0101] In some embodiments, digestion 206 further includes a quenching procedure. The quenching procedure may be performed by acidifying the sample (*e.g.*, to a pH <3). In some embodiments, formic acid may be used to perform this acidification.

[0102] In various embodiments, preparation workflow 200 further includes post-digestion procedure 207. Post-digestion procedure 207 may include, for example, a cleanup procedure. The cleanup procedure may include, for example, the removal of unwanted components in the sample that results from digestion 206. For example, unwanted components may include, but are not limited to, inorganic ions, surfactants, *etc.* In some embodiments, post-digestion procedure 207 further includes a procedure for the addition of heavy-labeled peptide internal standards.

[0103] Although preparation workflow 200 has been described with respect to a sample created or taken from biological sample 112 that is blood-based (*e.g.*, a whole blood sample, a plasma sample, a serum sample, *etc.*), sample preparation workflow 200 may be similarly implemented for other types of samples (*e.g.*, tears, urine, tissue, interstitial fluids, sputum, *etc.*) to produce set of peptides structures 122.

IV.B. Peptide Structure Identification and Quantitation

[0104] Figure 2B is a schematic diagram of data acquisition 124 in accordance with one or more embodiments. In various embodiments, data acquisition 124 can commence following sample preparation 200 described in Figure 2A. In various embodiments, data acquisition 124 can comprise quantification 208, quality control 210, and peak integration and normalization 212.

[0105] In various embodiments, targeted quantification 208 of peptides and glycopeptides can incorporate use of liquid chromatography-mass spectrometry LC/MS instrumentation. For example, LC-MS/MS, or tandem MS may be used. In general, LC/MS (*e.g.*, LC-MS/MS) can combine the physical separation capabilities of liquid chromatograph (LC) with the mass analysis capabilities of mass spectrometry (MS). According to some embodiments described herein, this technique allows for the separation of digested peptides to be fed from the LC column into the MS ion source through an interface.

[0106] In various embodiments, any LC/MS device can be incorporated into the workflow described herein. In various embodiments, a Triple Quadrupole LC/MSTM includes example instruments suited for identification and targeted quantification 208. In various embodiments, targeted quantification 208 is performed using multiple reaction monitoring mass spectrometry (MRM-MS).

[0107] In various embodiments described herein, identification of a particular protein or peptide and an associated quantity can be assessed. In various embodiments described herein, identification of a particular glycan and an associated quantity can be assessed. In various embodiments described herein, particular glycans can be matched to a glycosylation site on a protein or peptide and the abundance values measured.

[0108] In some cases, targeted quantification 208 includes using a specific collision energy associated for the appropriate fragmentation to consistently see an abundant product ion. Glycopeptide structures may have a lower collision energy than aglycosylated peptide

structures. When analyzing a sample that includes glycopeptide structures, the source voltage and gas temperature may be lowered as compared to generic proteomic analysis.

[0109] In various embodiments, quality control 210 procedures can be put in place to optimize data quality. In various embodiments, measures can be put in place allowing only errors within acceptable ranges outside of an expected value. In various embodiments, employing statistical models (*e.g.* using Westgard rules) can assist in quality control 210. For example, quality control 210 may include, for example, assessing the retention time and abundance of representative peptide structures (*e.g.*, glycosylated and/or aglycosylated) and spiked-in internal standards, in either every sample, or in each quality control sample (*e.g.*, pooled serum digest).

[0110] Peak integration and normalization 212 may be performed to process the data that has been generated and transform the data into a format for analysis. For example, peak integration and normalization 212 may include converting abundance data for various product ions that were detected for a selected peptide structure into a single quantification metric (*e.g.*, a relative quantity, an adjusted quantity, a normalized quantity, a relative concentration, an adjusted concentration, a normalized concentration, *etc.*) for that peptide structure. In some embodiments, peak integration and normalization 212 may be performed using one or more of the techniques described in U.S. Patent Publication No. 2020/0372973A1 and/or US Patent Publication No. 2020/0240996A1, the disclosures of which are incorporated by reference herein in their entireties.

V. Peptide Structure Data Analysis

V.A. Exemplary System for Peptide Structure Data Analysis

V.A.1. Analysis System for Peptide Structure Data Analysis

[0111] Figure 3 is a block diagram of an analysis system 300 in accordance with one or more embodiments. Analysis system 300 may be used, for example, to both detect and analyze various peptide structures that have been associated to various disease states. Analysis system 300 is one example of an implementation for a system that may be used to perform data analysis 108 in Figure 1. Thus, analysis system 300 is described with continuing reference to workflow 100 as described in Figures 1, 2A, and/or 2B.

[0112] Analysis system 300 may include computing platform 302 and data store 304. In some embodiments, analysis system 300 also includes display system 306. Computing platform 302 may take various forms. In one or more embodiments, computing platform 302 includes a

single computer (or computer system) or multiple computers in communication with each other. In other examples, computing platform 302 takes the form of a cloud computing platform.

5 [0113] Data store 304 and display system 306 may each be in communication with computing platform 302. In some examples, data store 304, display system 306, or both may be considered part of or otherwise integrated with computing platform 302. Thus, in some examples, computing platform 302, data store 304, and display system 306 may be separate components in communication with each other, but in other examples, some combination of these components may be integrated together. Communication between these different
10 components may be implemented using any number of wired communications links, wireless communications links, optical communications links, or a combination thereof.

[0114] Analysis system 300 includes, for example, peptide structure analyzer 308, which may be implemented using hardware, software, firmware, or a combination thereof. In one or more embodiments, peptide structure analyzer 308 is implemented using computing platform
15 302.

[0115] Peptide structure analyzer 308 receives peptide structure data 310 (*e.g.* abundance values for a transition) for processing. Peptide structure data 310 may be, for example, the peptide structure data that is output from sample preparation and processing 106 in Figures 1, 2A, and 2B. Peptide structure data 310 may correspond to set of peptide structures 122
20 identified for biological sample 112 and may thereby correspond to biological sample 112. In some embodiments, biological sample 112 can comprise a biological sample obtained from a subject. In some embodiments, biological sample 112 can comprise an internal standard. In some embodiments, biological sample 112 can comprise an external standard.

[0116] Peptide structure data 310 may comprise a set of quantification values for each
25 peptide structure of a plurality of peptide structures. In various embodiments, quantification values may comprise raw data received from a mass spectrometry instrument. Raw data may comprise a mass to charge ratio identifying a peptide structure and a raw abundance for an analyte (*e.g.* a peptide structure). In various embodiments, relative abundance can be determined from raw abundances. Quantification data for a peptide structures take the form of
30 any of the following prior to, during, and post analysis from one of a relative quantity, an adjusted quantity, and a normalized quantity. In this manner, peptide structure data 310 may provide abundance information about the plurality of peptide structures with respect to biological sample 112.

[0117] In some embodiments, a peptide structure of set of peptide structures comprises a glycosylated peptide structure, or glycopeptide structure, that is defined by a peptide sequence and a glycan structure attached to a linking site of the peptide sequence quantity. For example, the peptide structure may be a glycopeptide or a portion of a glycopeptide. In some
5 embodiments, a peptide structure of set of peptide structures 318 comprises an aglycosylated or non-glycosylated peptide structure that is defined by a peptide sequence. For example, the peptide structure may be a peptide or a portion of a peptide and may be referred to as a quantification peptide.

[0118] Peptide structure data 310 can be sent as input into peptide structure analyzer 308,
10 retrieved from data store 304 or some other type of storage (*e.g.*, cloud storage), accessed from cloud storage, or obtained in some other manner. In some cases, peptide structure data 310 may be retrieved from data store 304 in response to (*e.g.*, directly or indirectly based on) receiving user input entered by a user *via* an input device.

[0119] Peptide structure analyzer 308 may receive peptide structure data 310 for
15 processing. Received peptide structure data 310 by the peptide structure analyzer 308 may comprise raw abundance data from peptide structures (*e.g.*, raw abundance data for glycopeptide analytes, glycopeptide target analytes external standards, and/or internal standards). As such, peptide structure analyzer 308 may further process peptide structure data 310 using, for example, a normalization module for performing normalization 312 to generate
20 normalized peptide structure data. Normalization 312 may be implemented in any of a number of different ways and may include number of or combination of steps. Normalization 312 may be implemented using any number of models, functions, equations, algorithms, and/or other mathematical techniques. The present disclosure comprises novel systems and methods for normalizing peptide structure data 310 within an individual sample and across experiments.

[0120] In one or more embodiments, normalization 312 includes various steps such as, for
25 example, normalize data using external standard 362, normalize data using internal standard 364, relativistic normalization 366, and concentration determination 368. Aspects of normalization 312 can be carried out by inputting peptide structure data 310 into peptide structure analyzer 308. Normalization 312 of peptide structure analyzer 308 can process
30 peptide structure data using steps 362, 364, and 366. Processed peptide structure data can be used as input for concentration determination 368. In various embodiments, concentration determination 368 occurs as a last step.

[0121] In various aspects, the processes in normalization 312 can be executed to generate normalized relative abundance peptide structure data 324 from the peptide structure data 310. Peptide structure analyzer 308 may generate final output 128 based on normalized relative abundance peptide structure data 324. In other embodiments, final output 128 may be an output including normalized abundance data, normalized relative abundance peptide structure data, and/or may include data that has undergone model data processing 326. In other embodiments, final output 128 may be an output generated by normalization 312 and model data processing 326. Model data processing 326 may be performed to train a model and/or use a trained model for diagnosing a disease state.

[0122] For example, the normalized relative abundance peptide structure data 324 can be generated in a form suitable for model data processing 326. In some embodiments, normalized relative abundance peptide structure data 324 may include concentration values for peptide structures (*e.g.* target glycopeptide analyte concentrations) that can be comparable with respect to concentrations across multiple experiments.

[0123] In various embodiments, model data processing 326 may include, but is not limited to, at least one of a parametric model, a non-parametric model, deep learning model, a neural network, a linear discriminant analysis model, a quadratic discriminant analysis model, a support vector machine, a random forest algorithm, a nearest neighbor algorithm (*e.g.*, a k-Nearest Neighbors algorithm), a combined discriminant analysis model, a k-means clustering algorithm, an unsupervised model, a multivariable regression model, a penalized multivariable regression model, or another type of model.

[0124] Aspects of training predictive models, such as the ones listed above, may rely heavily on inputting large quantities of high-quality training data. In order to use large quantities of data, the data must first be made comparable across samples both within the same experiment and across different experiments. This can include accounting for contributors of variation in sample preparation, analysis, experimental conditions, and changes of conditions between experiment runs. Making and improving data comparison can be done using the systems and methods described herein and throughout.

V.A.2. Computer Implemented System

[0125] Figure 4 is a block diagram of a computer system in accordance with various embodiments. Computer system 400 may be an example of one implementation for computing platform 302 described above in Figure 3 and/or one or more systems in Figure 14.

[0126] In one or more examples, computer system 400 can include a bus 402 or other communication mechanism for communicating information, and a processor 404 coupled with bus 402 for processing information. In various embodiments, computer system 400 can also include a memory, which can be a random-access memory (RAM) 406 or other dynamic storage device, coupled to bus 402 for determining instructions to be executed by processor 404. Memory also can be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor 404. In various embodiments, computer system 400 can further include a read only memory (ROM) 408 or other static storage device coupled to bus 402 for storing static information and instructions for processor 404. A storage device 410, such as a magnetic disk or optical disk, can be provided and coupled to bus 402 for storing information and instructions.

[0127] In various embodiments, computer system 400 can be coupled *via* bus 402 to a display 412, such as a cathode ray tube (CRT) or liquid crystal display (LCD), for displaying information to a computer user. An input device 414, including alphanumeric and other keys, can be coupled to bus 402 for communicating information and command selections to processor 404. Another type of user input device is a cursor control 416, such as a mouse, a joystick, a trackball, a gesture input device, a gaze-based input device, or cursor direction keys for communicating direction information and command selections to processor 404 and for controlling cursor movement on display 412. This input device 414 typically has two degrees of freedom in two axes, a first axis (*e.g.*, x) and a second axis (*e.g.*, y), that allows the device to specify positions in a plane. However, it should be understood that input devices 414 allowing for three-dimensional (*e.g.*, x, y, and z) cursor movement are also contemplated herein.

[0128] Consistent with certain implementations of the present teachings, results can be provided by computer system 400 in response to processor 404 executing one or more sequences of one or more instructions contained in RAM 406. Such instructions can be read into RAM 406 from another computer-readable medium or computer-readable storage medium, such as storage device 410. Execution of the sequences of instructions contained in RAM 406 can cause processor 404 to perform the processes described herein. Alternatively, hard-wired circuitry can be used in place of or in combination with software instructions to implement the present teachings. Thus, implementations of the present teachings are not limited to any specific combination of hardware circuitry and software.

[0129] The term “computer-readable medium” (*e.g.*, data store, data storage, storage device, data storage device, *etc.*) or “computer-readable storage medium” as used herein refers to any media that participates in providing instructions to processor 404 for execution. Such a medium can take many forms, including but not limited to, non-volatile media, volatile media,
5 and transmission media. Examples of non-volatile media can include, but are not limited to, optical, solid state, magnetic disks, such as storage device 410. Examples of volatile media can include, but are not limited to, dynamic memory, such as RAM 406. Examples of transmission media can include, but are not limited to, coaxial cables, copper wire, and fiber optics, including the wires that comprise bus 402.

10 [0130] Common forms of computer-readable media include, for example, a floppy disk, a flexible disk, hard disk, magnetic tape, or any other magnetic medium, a CD-ROM, any other optical medium, punch cards, paper tape, any other physical medium with patterns of holes, a RAM, PROM, and EPROM, a FLASH-EPROM, any other memory chip or cartridge, or any other tangible medium from which a computer can read.

15 [0131] In addition to computer readable medium, instructions or data can be provided as signals on transmission media included in a communications apparatus or system to provide sequences of one or more instructions to processor 404 of computer system 400 for execution. For example, a communication apparatus may include a transceiver having signals indicative of instructions and data. The instructions and data are configured to cause one or more
20 processors to implement the functions outlined in the disclosure herein. Representative examples of data communications transmission connections can include, but are not limited to, telephone modem connections, wide area networks (WAN), local area networks (LAN), infrared data connections, NFC connections, optical communications connections, *etc.*

[0132] It should be appreciated that the methodologies described herein, flow charts, diagrams, and accompanying disclosure can be implemented using computer system 400 as a
25 standalone device or on a distributed network of shared computer processing resources such as a cloud computing network.

[0133] The methodologies described herein may be implemented by various means depending upon the application. For example, these methodologies may be implemented in
30 hardware, firmware, software, or any combination thereof. For a hardware implementation, the processing unit may be implemented within one or more application specific integrated circuits (ASICs), digital signal processors (DSPs), digital signal processing devices (DSPDs), programmable logic devices (PLDs), field programmable gate arrays (FPGAs), processors,

controllers, micro-controllers, microprocessors, electronic devices, other electronic units designed to perform the functions described herein, or a combination thereof.

[0134] In various embodiments, the methods of the present teachings may be implemented as firmware and/or a software program and applications written in conventional programming languages such as C, C++, Python, *etc.* If implemented as firmware and/or software, the embodiments described herein can be implemented on a non-transitory computer-readable medium in which a program is stored for causing a computer to perform the methods described above. It should be understood that the various engines described herein can be provided on a computer system, such as computer system 400, whereby processor 404 would execute the analyses and determinations provided by these engines, subject to instructions provided by any one of, or a combination of, the memory components RAM 406, ROM, 408, or storage device 410 and user input provided *via* input device 414.

VI. Exemplary Normalization Approaches

[0135] In various aspects of data collection, glycopeptide abundance readouts in targeted LC-MS experiments vary from run to run or experiment to experiment, or even for the same sample, based on digestion efficiency, ionization efficiency, detector sensitivity, run order effects, and other issues. The normalization methods below and throughout can correct for these and other issues.

[0136] Aspects of the disclosure include systems and methods for comparing data across multiple samples and experiments. In various embodiments, an initial step can include determining a corrected abundance using external standards. In various embodiments, another step can include determining non-glycosylated peptide concentrations using internal standards. In various embodiments, another step can include determining normalized/relative abundance of a glycopeptide analyte. In various embodiments, a final step can include determining concentration of a peptide analyte by multiplying the normalized/relative abundance value by the non-glycosylated endogenous peptide concentration. In various embodiments, the determining concentration step must occur subsequently to the previously listed steps.

VI.A. Normalize Data Using External Standard

[0137] Aspects of the disclosure can comprise determining for a corrected abundance for an individual transition (*e.g.* a non-glycosylated endogenous peptide [NGEP]) using a single-point calibration to an external standard.

[0138] Non-limiting examples of external standards can comprise pooled serum sample digest (e.g. pooled serum digest if biological samples to be analyzed is serum; pooled plasma digest if biological samples to be analyzed is plasma).

[0139] In various embodiments, when using pooled sample serum (e.g. Sigma Serum), markers can be selected that can be known to be present in the serum reference standard and the abundance differences can be used to align all abundance values of that marker in individual runs. In various embodiments, this can be done for any number of transitions. In some embodiments, there can be one transition per biomarker. In other embodiments, there can be one or more transitions per biomarker.

[0140] In various embodiments, two flanking pooled serum standard relative abundances can be used for a given marker (e.g. glycopeptide analyte).

[0141] Referring to Figure 8A, hypothetical abundance values for three experiments are shown graphically prior to applying the normalize data using external standards methods described above and herein according to various embodiments. Referring to Figure 8B, hypothetical abundance values for three experiments are shown graphically after applying the normalization methods described below and herein according to various embodiments. As illustrated, experimental data can be processed so as to allow for better comparisons across experiments in accordance with various embodiments.

VI.B. Normalize Data Using Internal Standard

[0142] In various embodiments, the following determination can be used in conjunction with use of internal standards. Internal standards can be useful for normalizing abundance values for a given sample. In some embodiments, internal standards can be “spiked-in” to each sample at specified, known concentrations and volumes. Internal standards can be effective in cases where enough is used to rise above a detection threshold (e.g. they can be detected).

Non-glycosylated endogenous peptide (NGEP) concentration = (corrected non-glycosylated endogenous peptide abundance / measured internal standard abundance) * known internal standard concentration.

[0143] In various embodiments, an NGEP concentration can later be used in calculating a concentration of a glycopeptide analyte. At least one internal standard per protein being analyzed can be measured in accordance with various embodiments.

[0144] In various embodiments, an internal standard can comprise a known concentration of a synthetic peptide corresponding to a given protein. In some embodiments, use of the

synthetic non-glycosylated peptide standards can generate a cost savings because glycosylated standards can be both difficult to come by and can be costly. In various embodiments, internal standards can comprise heavy-labeled peptides of a glycopeptide analyte target of an MRM-MS list. Alternatively, internal standards can comprise non-heavy labeled standards.

5 [0145] In some embodiments, internal standards can comprise fewer peptides than are being analyzed. Therefore, if an internal standard peptide is not available for each peptide for each glycoprotein, other internal standards can be selected as a “surrogate” internal standard based on similarity. In some embodiments, similarity comprises an m/z comparison. In other embodiments, similarity comprises an RT comparison.

10 VI.C. *Relativistic Normalization*

[0146] Aspects of the disclosure can be directed toward measuring an abundance of a glycopeptide analyte and determining a frequency of the glycopeptide analyte in accordance with various embodiments. In some embodiments, the frequency can comprise a relativistic normalization value. In some embodiments, the relativistic normalization value can include the
15 ratio of a measured abundance of a glycopeptide analyte (*e.g.* a glycoform) divided by a normalization factor (*e.g.* peptide abundance or a site occupancy total). In some embodiments, the determination can be based on making a comparison of the glycopeptide analyte to a non-glycosylated peptide from the same protein (*e.g.* peptide abundance comparison). In some embodiments, the determination can be based on making a comparison of the glycopeptide
20 analyte abundance to an abundance of other glycoforms at the same glycosylation site (*e.g.* a site occupancy total).

[0147] In various embodiments, using a relative abundance approach (*e.g.* a comparison of the glycopeptide analyte to a non-glycosylated peptide from the same protein) to relativistic normalization may be used when two or less glycoforms exist for a given glycosylation site. In
25 various embodiments, using a site occupancy abundance approach (*e.g.* a comparison of the glycopeptide analyte abundance to an abundance of other glycoforms at the same glycosylation site) to relativistic normalization may be used when three or more glycoforms exist for a given glycosylation site.

[0148] In various embodiments, relativistic normalization can be or produce a value
30 comprising information relating to what fraction a particular glycoform (*e.g.* glycopeptide analyte) occurs relative to a total of all glycoforms for a given glycosylation site. In some embodiments, the value can comprise a percentage. In some embodiments, the value can

comprise a fraction. In some embodiments, the value can comprise an abundance value for a glycopeptide analyte accounting for abundance values of all glycoforms of a peptide.

VI.C.1.Relative Abundance

[0149] Aspects of the disclosure consider and correct for variation generated within a single experiment such as well-to-well variation, patient to patient, and run order.

[0150] In various aspects, normalized-relative abundance of a glycopeptide analyte can be determined by dividing the measured abundance of the target glycopeptide analyte (*e.g.* a glycoform) by the measured abundance of a non-glycosylated peptide of the same protein. As previously stated, a relative abundance can be calculated as a measured abundance value for a glycoform (*e.g.* a glycopeptide or a marker) divided by a corresponding quantification peptide abundance for various embodiments. For a hypothetical example, (glycopeptide abundance 3000)/(quantification peptide abundance for another non-glycosylated peptide of the same protein 30000) = 10% glycosylation.

[0151] In some embodiments, an advantage of the relative abundance approach can be that it may account for any kind of glycopeptide no matter how many glycans/glycan sites exist or are measured for a given glycoprotein or peptide.

VI.C.2.Site Occupancy Abundance

[0152] In various embodiments, a site occupancy approach can be used in determining relativistic abundance. In various embodiments, a site occupancy approach can comprise determining a per-site glycopeptide occupancy proportion across all glycan observed at that site. In some embodiments, such an approach can be more effective when more than one glycoform is present on site. In some embodiments, abundance values can be measured for each glycoform and then totaled. A relativistic abundance value can then be determined by dividing an abundance value for a single glycoform by the total abundances of all glycoforms for a given glycosylation site. For a hypothetical example, (glycoform 1 abundance) can be divided by (glycoform 1 abundance + glycoform 2 abundance + glycoform 3 abundance)

[0153] In various embodiments, the site occupancy approach can account for ionization efficiencies across glycopeptides.

[0154] In various embodiments, instead of using site occupancy for a given protein, a sum of abundances of all quantification peptides for a given subject can be used.

VI.D. Concentration Determination

[0155] In various embodiments, a final step can be completed to obtain an approximate concentration of a glycopeptide analyte. In various embodiments an approximate concentration of a glycopeptide analyte can be determined by multiplying the normalized (relative) abundance value or site occupancy abundance value, as appropriately obtained above, by the non-glycosylated endogenous peptide concentration obtained above.

[0156] In various embodiments, a process for concentration determination can improve normalization results when combined with the other processes described here. Please see *X. Experimental Results* and Figures 7A and 7B.

VII. Exemplary Normalization Methods

VII.A. Normalized Abundance Methods

[0157] In various embodiments, any of the methods described herein can be performed on any system capable of carrying out the describes individual steps of the method. A non-limiting example of a system is described in Figure 3 along with associated description provided herein.

[0158] In various embodiments, the below steps can be stored on a non-transitory computer-readable medium including computer instructions that, when executed by a computer, cause the computer to perform the below method for determining a concentration of a target glycopeptide analyte.

[0159] Referring to Figure 5, a method of determining a concentration of a target glycopeptide analyte in a sample is described according to some embodiments.

[0160] Step 502 comprises measuring a non-glycosylated endogenous peptide (NGEP) abundance value for an NGEP in a sample in accordance with various embodiments.

[0161] Step 504 comprises determining a corrected NGEP abundance value for the transition using an external standard in accordance with various embodiments.

[0162] Step 506 comprises measuring an internal standard abundance for an internal standard having a known internal standard concentration in accordance with various embodiments.

[0163] Step 508 comprises determining a NGEP concentration value as a function of the corrected NGEP abundance value, the known internal standard concentration, and the internal standard abundance in accordance with various embodiments.

[0164] Step 510 comprises determining a normalized abundance value of a target glycopeptide analyte as a function of a first measured abundance of the target glycopeptide

analyte and a second measured abundance of a non-glycosylated peptide on the same protein as the target glycoprotein analyte in accordance with various embodiments.

5 [0165] Step 512 comprises determining a target glycopeptide analyte concentration as a function of the NGEP concentration value and the normalized abundance value in accordance with various embodiments.

[0166] In various embodiments, an NGEP concentration value comprises a ratio of the corrected NGEP abundance value and the internal standard abundance. In various embodiments, an NGEP concentration value can be a product of the ratio and the known internal standard concentration.

10 [0167] In various embodiments, a normalized abundance value can be determined when one or two glycans are present at a target glycosylation site on the protein.

[0168] In various embodiments, a target glycopeptide analyte concentration can be a product of an NGEP concentration value and a normalized abundance value.

15 [0169] In various embodiments, a normalized abundance value can be the quotient of a first measured abundance and a second measured abundance.

[0170] In various embodiments, the abundance for the NGEP, the internal standard abundance, and the measured abundance of the target glycopeptide can be measured in a sample run using mass spectrometry.

VII.B. Site Occupancy Methods

20 [0171] In various embodiments, any of the methods described herein can be performed on any system capable of carrying out the describes individual steps of the method. A non-limiting example of a system is described in Figure 3 along with associated description provided herein.

[0172] In various embodiments, the below steps can be stored on a non-transitory computer-readable medium including computer instructions that, when executed by a
25 computer, cause the computer to perform the below method for determining a concentration of a target glycopeptide analyte in a sample.

[0173] Referring to Figure 6, a method of determining a concentration of a target glycopeptide analyte in a sample is described in accordance with various embodiments.

30 [0174] Step 602 comprises obtaining a non-glycosylated endogenous peptide (NGEP) abundance value for a NGEP in a sample in accordance with various embodiments.

[0175] Step 604 comprises determining a corrected NGEP abundance value for the transition using an external standard in accordance with various embodiments.

[0176] Step 606 comprises measuring an internal standard abundance for an internal standard having a known internal standard concentration in accordance with various embodiments.

5 [0177] Step 608 comprises determining a NGEP concentration value as a function of the corrected NGEP abundance value, the known internal standard concentration, and the internal standard abundance in accordance with various embodiments.

[0178] Step 610 comprises determining a site occupancy value of a target glycopeptide analyte as a function of a measured target abundance of the target glycopeptide analyte at a given site, and a measured total abundance of all glycopeptides at the same site in accordance
10 with various embodiments.

[0179] Step 612 comprises determining a target glycopeptide analyte concentration as a function of the NGEP concentration value and the normalized abundance value in accordance with various embodiments.

[0180] In various embodiments, an NGEP concentration value comprises a ratio of the
15 corrected NGEP abundance value and the internal standard abundance. In various embodiments, an NGEP concentration value can be a product of the ratio and the known internal standard concentration.

[0181] In various embodiments, a normalized abundance value can be determined when one or two glycans are present at a target glycosylation site on the protein.

20 [0182] In various embodiments, a target glycopeptide analyte concentration can be a product of an NGEP concentration value and a normalized abundance value.

[0183] In various embodiments, a normalized abundance value can be the quotient of a first measured abundance and a second measured abundance.

[0184] In various embodiments, the abundance for the NGEP, the internal standard
25 abundance, and the measured abundance of the target glycopeptide can be measured in a sample run using mass spectrometry.

VII.C. Exemplary Methods of Determining Glycopeptide Concentration

[0185] In various embodiments, any of the methods described herein can be performed on any system capable of carrying out the describes individual steps of the method. A non-limiting
30 example of a system is described in Figure 3 along with associated description provided herein.

[0186] In various embodiments, the below steps can be stored on a non-transitory computer-readable medium including computer instructions that, when executed by a

computer, cause the computer to perform the below method for determining a concentration of a target glycopeptide analyte in a sample.

[0187] Referring to Figure 9, a method of determining a concentration of a target glycopeptide analyte in a sample is described in accordance with various embodiments.

5 [0188] Step 902 comprises receiving raw abundance data for a sample from a mass spectrometry system, the raw abundance data comprising a raw external standard abundance for an external standard in the sample, a raw internal standard abundance for an internal standard that has a known concentration in the sample, a raw glycopeptide abundance for a glycopeptide structure in the sample, a raw non-glycosylated peptide abundance for a non-glycosylated peptide structure in the sample, the non-glycosylated peptide structure being
10 derived from a same glycoprotein as the glycopeptide structure. In various embodiments, a single run can analyze a sample comprising an external standard, an internal standard, and a glycopeptide analyte. As such, abundance values (*e.g.* abundance or raw abundance) for external standard, internal standard, and target glycopeptide analyte can be determined by mass
15 spectrometry in the same run.

[0189] Step 904 comprises computing corrected abundance data using a difference between the raw external standard abundance and a reference abundance for the external standard. In some embodiments, computing the corrected abundance data comprises computing a corrected abundance for the non-glycosylated peptide structure.

20 [0190] Step 906 comprises generating normalized concentration data for the sample using the corrected abundance data, the raw internal standard abundance, the known concentration of the internal standard, the raw glycopeptide abundance, and the raw non-glycosylated peptide abundance. In some embodiments, generating comprises computing a non-glycosylated peptide concentration using the corrected abundance and the raw internal standard abundance.
25 In some embodiments, generating comprises computing a normalized abundance for the glycopeptide structure using the raw glycopeptide abundance for the glycopeptide structure and the raw non-glycosylated peptide abundance for the non-glycosylated peptide structure. In some embodiments, generating comprises computing a normalized concentration for the glycopeptide structure using the normalized abundance for the glycopeptide structure and the
30 non-glycosylated peptide concentration for the non-glycosylated peptide structure, wherein the concentration data includes the normalized concentration for the glycopeptide structure.

[0191] In various embodiments, the normalized concentration data can be normalized to account for variances across a plurality of experiments run on the mass spectrometry system.

[0192] In various embodiments, the normalized concentration data can be normalized to account for variances across a plurality of experiments run on different mass spectrometry systems.

5 [0193] In various embodiments, the normalized concentration data can be normalized to account for variances across a plurality of experiments run in different laboratories.

[0194] In various embodiments, the normalized concentration data can be normalized to account for variances when analyzing a plurality of samples across at least one of a plurality of experiments, a plurality of laboratories, or a plurality of different mass spectrometry systems.

10 [0195] Step 908 comprises analyzing the normalized concentration data using a model system to generate an output for a subject. In various embodiments, the output comprises a treatment output. In various embodiments, the output comprises a diagnosis output.

VIII. Exemplary Sample Preparation and LC-MS Run Order Methods

15 [0196] Predictive capabilities of predictive algorithms can be influenced by sample preparation and LC-MS run order methods. The methods and systems for data normalization described herein can be influenced by the individual constituents (*e.g.*, a patient sample and one or more standards) of the run samples and the position of the run sample on a run order.

[0197] Figure 13 illustrates a flowchart of an exemplary method for preparing samples for a LC-MS sample run and normalizing abundance data of a plurality of sample runs 1300.

20 [0198] Step 1302 includes preparing run samples for a plurality of sample runs for LC-MS analysis. In many methods, preparing run samples can include generating a first set of run samples. In many methods, preparing run samples can include generating a second set of run samples.

25 [0199] Each of the first set of run samples can include an external standard. In many embodiments, the external standard can originate from an external standard source. In some embodiments, the external standard source can include a serum. In various embodiments, an external standard can include an NGEF. In various embodiments, a source for an external standard can match a source for a biological sample. In various embodiments, a source includes plasma. In some embodiments, a source includes whole blood. In some embodiments, a source includes tissue. A non-limiting example may include pooled plasma digests being used as
30 external standards when biological/patient samples to be characterized include plasma samples.

[0200] In various embodiments, a pooled standard originating from a serum source (*e.g.*, sigma serum or pooled plasma serum), being used as an external standard in the method of

normalization described herein, does not have a known concentration. (*e.g.*, a constituent analyte prior to the LC/MS). In various embodiments, the concentration of the external standard may be unknown prior to LC-MS analysis. In various embodiments, the concentration of the pooled standard may be unknown prior to LC-MS analysis.

5 **[0201]** In various embodiments, a type of an external standard can match a type of a biological sample. In various embodiments, a type can include a peptide structure or sequence of a peptide structure (*e.g.*, a standard and a glycopeptide analyte may have a sequence similarity greater than 99%, 95%, 90%, or 85%). In various embodiments, a type can include a peptide structure including a glycosylation site. In various embodiments, a type can include
10 an m/z ratio (*e.g.*, a standard and a glycopeptide analyte having an m/z ratio similarity greater than 80%, 85%, 90%, 95%, or 99%). In various embodiments, a type can include an abundance (*e.g.*, a standard and a glycopeptide analyte having an abundance similarity greater than 80%, 85%, 90%, 95%, or 99%).

15 **[0202]** In certain embodiments, a type of biological sample used for an external standard may be matched to the source or type of biological sample being characterized in a patient sample, the abundances generated from peptide structures being normalized (*e.g.* pooled plasma digests to plasma samples, pooled serum digests to serum samples, pooled whole blood digests to whole blood samples).

20 **[0203]** Each of the second set can include at least two target glycopeptide analytes and an internal standard.

[0204] The step of preparing can further comprise enzymatically digesting glycoprotein structures of the various run samples to generate the external standards, the internal standards, and the at least two target glycopeptide analytes of the run samples. Exemplary processing protocols including digestion are discussed through the present description.

25 **[0205]** Step 1304 includes combining at least two run samples of the first set to create a pooled standard.

30 **[0206]** Step 1306 includes analyzing each run sample according to a run order, wherein the run order specifies a relative order of analysis for the run samples of the sample run. A run sample of the first set can occupy an adjacent position in the run order to a run sample of the second set, in protocols. In many protocols, a run sample of the first set can occupy a position before all of the run samples of the second set. In additional and other protocols, a run sample of the first set can occupy a position after all of the run samples of the second set.

[0207] In some protocols, the run order comprises analyzing one run sample of the first set at least every five positions in the run order. In other protocols, the run order comprises analyzing one run sample of the first set at least every ten positions in the run order. In still other protocols, the run order comprises analyzing one run sample of the first set at least every
5 fifteen positions in the run order.

[0208] For most protocols, the run order comprises analyzing one sample of the first set followed by a range of between two and twenty-nine run samples of the second set. In many protocols, the run order comprises analyzing one sample of the first set followed by a range of between four and fourteen run samples of the second set.

10 [0209] Additionally, protocols can include runs orders that specify analyzing at least one run sample comprising BSA before all the run samples of the first and second sets have been analyzed. Still other protocols can include analyzing at least one run sample comprising BSA after all the run samples of the first and second sets have been analyzed.

[0210] Protocols can include analyzing at least one blank run sample prior to analyzing a
15 run sample of the first or the second set in a run order. Protocols for run order can include analyzing a blank run sample one position before analyzing a sample of the first set in the run order. In various embodiments, blank run samples can include water.

[0211] Step 1308 includes normalizing the raw abundance data of the plurality of sample runs by using the pooled standard of each sample run as a reference.

20 [0212] Some processes include normalizing raw abundance data generated from each of the plurality of target glycopeptide analytes of the same run sample using the internal standards. Internal standard can comprise a non-glycosylated endogenous peptide (NGEP). In some embodiments, each target glycopeptide analyte can have a corresponding internal standard for a reference. In some cases, an internal standard can comprise a surrogate.

25 IX. Exemplary Sample Processing System

[0213] Figure 14 illustrates an exemplary sample processing system 1400 for processing patient sample into normalized abundance. A sample processing system 1400 can include a sample preparation system 1402, a sample analysis system 1404, and a data analysis system 1406. The sample processing systems described herein can comprise hardware and instructions
30 for carrying out the processes described in Figures 1, 2A, 2B, 5, 6, 9, and 13.

[0214] A sample preparation system 1402 can include a fluidic instrument for processing glycoproteins into glycopeptide analytes in accordance with the various embodiments

described herein. The fluidic instrument can include one or more pipettes. The sample preparation system 1402 can include partitions for storing run samples prior to analysis, for mixing, and for other sample predation steps (*e.g.*, digestion). Non-limiting examples of partitions can include vials and wells (*e.g.*, wells of a 96-well plate).

- 5 **[0215]** The sample preparation system 1402 can generate a first set of run samples, each of the first set including an external standard. In various embodiment, an external standard can include an NGEP. In various embodiments, a source for an external standard can match a source for a biological sample. In various embodiments, a source includes plasma. In some embodiments, a source includes whole blood. In some embodiments, a source includes tissue.
- 10 A non-limiting example may include pooled plasma digests being used as external standards when biological/patient samples to be characterized include plasma samples. In various embodiments, a pooled standard originating from a serum source (*e.g.*, sigma serum or pooled plasma serum), being used as an external standard in the method of normalization described herein, does not have a known concentration. (*e.g.*, a constituent analyte prior to the LC/MS).
- 15 In various embodiments, the concentration of the external standard may be unknown prior to LC-MS analysis. In various embodiments, the concentration of the pooled standard may be unknown prior to LC-MS analysis.

- [0216]** In various embodiments, a type of an external standard can match a type of a biological sample. In various embodiments, a type can include a peptide structure or sequence
- 20 of a peptide structure (*e.g.*, a standard and a glycopeptide analyte may have a sequence similarity greater than 99%, 95%, 90%, or 85%). In various embodiments, a type can include a peptide structure including a glycosylation site. In various embodiments, a type can include an m/z ratio (*e.g.*, a standard and a glycopeptide analyte having an m/z ratio similarity greater than 80%, 85%, 90%, 95%, or 99%). In various embodiments, a type can include an abundance
- 25 (*e.g.*, a standard and a glycopeptide analyte having an abundance similarity greater than 80%, 85%, 90%, 95%, or 99%).

- [0217]** In certain embodiments, a type of biological sample used for an external standard may be matched to the source or type of biological sample being characterized in a patient sample, the abundances generated from peptide structures being normalized (*e.g.* pooled
- 30 plasma digests to plasma samples, pooled serum digests to serum samples, pooled whole blood digests to whole blood samples).

[0218] The sample preparation system 1402 can generate a second set of run samples, each of the second set including at least two target glycopeptide analytes and an internal standard,

The sample preparation system 1402 can combine at least two run samples of the first set to create a pooled standard. The combining step can occur after enzymatic digestion in some embodiments. In other embodiments, the combining step can occur before enzymatic digestion. The step of preparing can further comprise enzymatically digesting glycoprotein structures of the run samples to generate the external standards, the internal standards, and the at least two target glycopeptide analytes of the run samples. Non-limiting examples of preparation processes that can be carried out on the sample preparation system 1402 are illustrated in Figure 2A and described in Section IV.A. Sample Preparation and Processing.

[0219] The sample analysis system 1404 can include a LC-MS instrument for analyzing glycopeptide analytes. A process comprising analyzing each run sample according to a run order can be carried out on the LC-MS, wherein the run order can be stored on a data store in electronical communication with the LC-MS instrument and specifies a relative order of analysis for the run samples of the sample run. Non-limiting examples of sample analysis that can be carried out on the sample analysis system 1404 are illustrated in Figure 2B and described in Section IV.B. Peptide Structure Identification and Processing.

[0220] In many processes a run sample of the first set occupies an adjacent position in the run order to a run sample of the second set. In many processes a run sample of the first set occupies a position in the run order before all of the run samples of the second set. In many processes a run sample of the first set occupies a position in the run order after all of the run samples of the second set. In some processes, the run order comprises analyzing one run sample of the first set at least every five positions in the run order. In more processes, the run order comprises analyzing one run sample of the first set at least every ten positions in the run order. In many processes, the run order comprises analyzing one run sample of the first set at least every fifteen positions in the run order. The run order can comprise analyzing one sample of the first set followed by a range of between two and twenty-nine run samples of the second set. The run order can comprise analyzing one sample of the first set followed by a range of between four and fourteen run samples of the second set. In some systems the process can include analyzing at least one run sample comprising BSA before all the run samples of the first and second sets have been analyzed. Some processes of the system include analyzing at least one run sample comprising BSA after all the run samples of the first and second sets have been analyzed. Some processes of the system include analyzing at least one blank run sample prior to analyzing a run sample of the first or the second set. Some processes of the system include

analyzing a blank run sample one position before analyzing a sample of the first set in the run order.

[0221] The data analysis system 1406 can further include a normalization module. Normalization 312 shown in Figure 3 can be an example of the normalization module. The
5 exemplary normalization methods described herein can be processes using the data normalization module (see Section VII. Exemplary Normalization Methods). The process further comprises normalizing raw abundance data generated from each of the plurality of target glycopeptide analytes of the same run sample using the internal standards.

[0222] In many systems the external standard comprises a non-glycosylated endogenous
10 peptide (NGEP). In many systems, each target glycopeptide analyte can have a corresponding internal standard for a reference. At least some of the internal standards can be surrogates. In many systems, serum comprises the external standard and can be an external standard source.

X. Experimental Results

X.A. Normalization Methods

15 [0223] The experimental results described below demonstrate the differential accuracy achieved using the systems and methods described herein versus alternatives.

[0224] Referring to Figure 7A, mean concentrations with 99% confidence intervals for six markers are presented. The mean concentrations for the six markers were determined using the following method.

20 [0225] In determining a mean concentration Figure 7A uses a process that employs no external standard-based abundance correction (see **VI.A. Normalize Data Using External Standard**), and this leads to misaligned results in peptide concentration across three separate runs, as observed in the difference between the HVI-CRC run and the SS-JAR1 run.

[0226] Referring to Figure 7B, mean concentrations for the six marks are presented when
25 completing an abundance correction step (*e.g.* Normalize Data Using External Standard) prior to the process used in generating the data shown in Figure 7A, namely, for every new run, on a per-marker basis, abundance data was corrected for differences in ionization and digestion efficiency by multiplying by the quotient of Sigma Serum means of the new run and some reference. $\text{Corrected_Patient_Abundance} = \text{Patient_Abundance} * (\text{ReferenceRun_Mean_Sigma_Serum_Abundance} / \text{NewRun_Mean_Sigma_Serum_Abundance})$. This results in
30 aligned peptide concentrations, which allows for comparing aligned approximate glycopeptide concentrations.

[0227] Figure 7A compared to Figure 7B demonstrate that without a specified order of steps (*e.g.* Normalize Data Using External Standard step occurring first and Concentration Determination occurring last), peptide level data will not align. As such, various embodiments described herein benefit in ordering the Normalize Data Using External Standard step first and
5 Concentration Determination last.

[0228] Figure 10A and Figure 10B illustrate a comparison of independent ovarian cancer and benign cohorts in batch 1 (“B1”) and batch 2 (“B2”). Figure 10A represents a single glycopeptide concentration calculated *via* site occupancy and the Normalize Data Using External Standard step occurring first and Concentration Determination occurring last,
10 resulting in similar disease distributions.

[0229] Figure 10B displays the same marker and patients without using the abundance correction to an external standard step (*e.g.* the Normalize Data Using External Standard step), resulting in batch 2 values that do not align with batch 1. As such, the data in Figure 10B cannot be effectively compared whereas the data in Figure 10A can be effectively compared and used
15 in conjunction with predictive models.

[0230] Figure 11A and Figure 11B illustrate a comparison of independent ovarian cancer and benign cohorts in batch 1 (“B1”) and batch 2 (“B2”). Figure 10A represents a single glycopeptide concentration calculated *via* relative abundance and the Normalize Data Using External Standard step occurring first and Concentration Determination occurring last,
20 resulting in similar disease distributions.

[0231] Figure 11B displays the same marker and patients without using the abundance correction to an external standard step (*e.g.* the Normalize Data Using External Standard step), resulting in batch 2 values that do not align with batch 1. As such, the data in Figure 11B cannot be effectively compared whereas the data in Figure 11A can be effectively compared and used
25 in conjunction with predictive models.

[0232] Figure 12A and Figure 12B illustrate confusion matrices resulting in applying a trained ovarian cancer model to an independent test set. Figure 12A represents a model where concentration was calculated using a Normalize Data Using External Standard step first and Concentration Determination last. Figure 12B represents a model where concentration was calculated omitting the Normalize Data Using External Standard step. Overall, the predictive accuracy seen in Figure 12A was determined to be 0.884 and the predictive accuracy seen in Figure 12B was determined to be 0.71. As such, completing Normalize Data Using External Standard step first, following by completing the Normalize Data Using Internal Standard and
30

Relativistic Normalization steps and finally the Concentration Determination was shown to increase predictive accuracy by 0.174.

X.B. Sample Preparation and Processing

[0233] Figure 15 illustrates experimental results for data including coefficient of variation for pooled replicates. The data demonstrates that including too few external standards (*e.g.*, sigma serum (SS) replicates) in an experiment may not allow accurate estimations for a coefficient of variation (CV) of biomarkers (*e.g.*, glycopeptide analytes) undergoing mass spectrometry analysis. The median CV of all biomarkers is plotted for each possibility from $n=2$ SS for the entire experiment (left line) up to one SS every 10 patient samples (right, with $n=33$ SS for $n=319$ patients). As the frequency of SS replicates increases, the variance around this median CV decreases to a low amount as the estimation approaches the true median CV (0.124).

[0234] Figure 16 illustrates an experimental run order for a set of run samples. Run samples include blanks, BSA, system buffer, target glycopeptide analytes, internal standards, and external standards. In the run order pooled standard, "SS-JAR1-SysSuit," is run every 10th position and a blank in the position just prior in a run order. A sample position coordinate system can be used for robot fluidic instrument mediated sample preparation. The run order was generated using a run order algorithm. The algorithm was written into a script that automatically generated a run sequence based on patient sample data. The algorithm took a list of patient samples with external sample IDs and their corresponding internal sample IDs, then randomized the samples and put a SS digest at the beginning and after every 10 randomized patient samples and assigned them sample preparation numbers starting from #1, and then the samples were randomized again and put into the designed run sequence pattern with interspersed blanks and SS pool runs.

[0235] Figure 17 illustrates a layout for a 96-well plate including run sample partition locations. Non-limiting constituents of run samples can include blanks, BSA, system buffer, target glycopeptide analytes, internal standards, and external standards. "ss" followed by a number identify locations of external standards, "SS" identifies pooled standards, the other markers indicate other run samples (*e.g.*, run samples including target glycopeptide analytes and/or internal standards).

XI. Additional Considerations

[0236] Any headers and/or subheaders between sections and subsections of this document are included solely for the purpose of improving readability and do not imply that features cannot be combined across sections and subsection. Accordingly, sections and subsections do
5 not describe separate embodiments.

[0237] While the present teachings are described in conjunction with various embodiments, it is not intended that the present teachings be limited to such embodiments. On the contrary, the present teachings encompass various alternatives, modifications, and equivalents, as will be appreciated by those of skill in the art. The present description provides preferred exemplary
10 embodiments, and is not intended to limit the scope, applicability or configuration of the disclosure. Rather, the present description of the preferred exemplary embodiments will provide those skilled in the art with an enabling description for implementing various embodiments.

[0238] It is understood that various changes may be made in the function and arrangement of elements without departing from the spirit and scope as set forth in the appended claims. Thus, such modifications and variations are considered to be within the scope set forth in the appended claims. Further, the terms and expressions which have been employed are used as terms of description and not of limitation, and there is no intention in the use of such terms and expressions of excluding any equivalents of the features shown and described or portions
20 thereof, but it is recognized that various modifications are possible within the scope of the invention claimed.

[0239] In describing the various embodiments, the specification may have presented a method and/or process as a particular sequence of steps. However, to the extent that the method or process does not rely on the particular order of steps set forth herein, the method or process
25 should not be limited to the particular sequence of steps described, and one skilled in the art can readily appreciate that the sequences may be varied and still remain within the spirit and scope of the various embodiments.

[0240] Some embodiments of the present disclosure include a system including one or more data processors. In some embodiments, the system includes a non-transitory computer
30 readable storage medium containing instructions which, when executed on the one or more data processors, cause the one or more data processors to perform part or all of one or more methods and/or part or all of one or more processes disclosed herein. Some embodiments of the present disclosure include a computer-program product tangibly embodied in a

non-transitory machine-readable storage medium, including instructions configured to cause one or more data processors to perform part or all of one or more methods and/or part or all of one or more processes disclosed herein.

[0241] Specific details are given in the present description to provide an understanding of the embodiments. However, it is understood that the embodiments may be practiced without these specific details. For example, circuits, systems, networks, processes, and other components may be shown as components in block diagram form in order not to obscure the embodiments in unnecessary detail. In other instances, well-known circuits, processes, algorithms, structures, and techniques may be shown without unnecessary detail in order to avoid obscuring the embodiments.

XII. Recitation of Embodiments

[0242] Embodiment 1: A method of determining a concentration of a target glycopeptide analyte in a sample, the method comprising: measuring an abundance for a non-glycosylated endogenous peptide (NGEP) of a sample; determining a corrected NGEP abundance for the NGEP using an external standard; measuring an internal standard abundance for an internal standard having a known internal standard concentration; determining an NGEP concentration for the NGEP as a function of the corrected NGEP abundance, the known internal standard concentration, and the internal standard abundance; determining a normalized abundance of a target glycopeptide analyte of the sample as a function of a measured abundance of the target glycopeptide analyte and the measured abundance of the NGEP on the same protein as the target glycoprotein analyte; and determining a target glycopeptide analyte concentration as a function of the NGEP concentration and the normalized abundance.

[0243] Embodiment 2: The method of embodiment 1, further comprising analyzing the target glycopeptide analyte concentration to generate a treatment for a subject.

[0244] Embodiment 3: The method of embodiments 1 or 2, further comprising analyzing the target glycopeptide analyte concentration to generate a diagnosis for a subject.

[0245] Embodiment 4: The method of any one of embodiments 1-3, wherein the NGEP concentration comprises a ratio of the corrected NGEP abundance and the internal standard abundance.

[0246] Embodiment 5: The method of embodiment 4, wherein the NGEP concentration is a product of the ratio and the known internal standard concentration.

[0247] Embodiment 6: The method of any one of embodiments 1-5, wherein the normalized abundance is determined when one or two glycans are identified at a target glycosylation site on the protein.

5 [0248] Embodiment 7: The method of any one of embodiments 1-6, wherein the target glycopeptide analyte concentration is a product of the NGEP concentration and the normalized abundance.

[0249] Embodiment 8: The method of any one of embodiments 1-7, wherein the normalized abundance is the quotient of the first measured abundance and the second measured abundance

10 [0250] Embodiment 9: The method of any one of embodiments 1-8, wherein the abundance for the NGEP, the internal standard abundance, and the measured abundance of the target glycopeptide are measured in a sample run using mass spectrometry.

[0251] Embodiment 10: A non-transitory computer-readable medium storing computer instructions that, when executed by a computer, cause the computer to perform a method for
15 determining a concentration of a target glycopeptide analyte in a sample, the method comprising: measuring an abundance for a non-glycosylated endogenous peptide (NGEP) of a sample; determining a corrected NGEP abundance for the NGEP using an external standard; measuring an internal standard abundance for an internal standard having a known internal standard concentration; determining an NGEP concentration as a function of the corrected
20 NGEP abundance, the known internal standard concentration, and the internal standard abundance; determining a normalized abundance of a target glycopeptide analyte of the sample as a function of a measured abundance of the target glycopeptide analyte and the measured abundance of the NGEP on the same protein as the target glycoprotein analyte; and determining a target glycopeptide analyte concentration as a function of the NGEP concentration and the
25 normalized abundance.

[0252] Embodiment 11: The method of embodiment 10, further comprising analyzing the target glycopeptide analyte concentration to generate a treatment for a subject.

[0253] Embodiment 12: The method of embodiments 10 or 11, further comprising analyzing the target glycopeptide analyte concentration to generate a diagnosis for a subject.

30 [0254] Embodiment 13: The method of any one of embodiments 10-12, wherein the NGEP concentration comprises a ratio of the corrected NGEP abundance and the internal standard abundance.

[0255] Embodiment 14: The method of embodiment 13, wherein the NGEP concentration is a product of the ratio and the known internal standard concentration.

[0256] Embodiment 15: The method of any one of embodiments 10-14, wherein the normalized abundance is determined when one or two glycans are identified at a target glycosylation site on the protein.

[0257] Embodiment 16: The method of any one of embodiments 10-15, wherein the target glycopeptide analyte concentration is a product of the NGEP concentration and the normalized abundance.

[0258] Embodiment 17: The method of any one of embodiments 10-16, wherein the normalized abundance is the quotient of the first measured abundance and the second measured abundance.

[0259] Embodiment 18: The method of any one of embodiments 10-17, wherein the abundance for the NGEP, the internal standard abundance, and the measured abundance of the target glycopeptide are measured in a sample run using mass spectrometry.

[0260] Embodiment 19: A method of determining a concentration of a target glycopeptide analyte in a sample, the method comprising: measuring an abundance for a non-glycosylated endogenous peptide (NGEP) of a sample; determining a corrected NGEP abundance for the NGEP using an external standard; measuring an internal standard abundance for an internal standard having a known internal standard concentration; determining a NGEP concentration as a function of the corrected NGEP abundance, the known internal standard concentration, and the internal standard abundance; determining a site occupancy of a target glycopeptide analyte of the sample as a function of a measured target abundance of the target glycopeptide analyte at a given site, and a measured total abundance of all glycopeptides quantified at the same site; and determining a target glycopeptide analyte concentration as a function of the NGEP concentration and the normalized abundance.

[0261] Embodiment 20: The method of embodiment 19, further comprising analyzing the target glycopeptide analyte concentration to generate a treatment for a subject.

[0262] Embodiment 21: The method of embodiments 19 or 20, further comprising analyzing the target glycopeptide analyte concentration to generate a diagnosis for a subject.

[0263] Embodiment 22: The method of any one of embodiments 19-21, wherein the NGEP concentration comprises a ratio of the corrected NGEP abundance and the internal standard abundance

[0264] Embodiment 23: The method of embodiment 22, wherein the NGEP concentration is a product of the ratio and the known internal standard concentration.

[0265] Embodiment 24: The method of any one of embodiments 19-23, wherein the site occupancy is determined when three or more glycans are identified at the given site.

5 [0266] Embodiment 25: The method of any one of embodiments 19-24, wherein the target glycopeptide analyte concentration is a product of the NGEP concentration and the normalized abundance.

[0267] Embodiment 26: The method of any one of embodiments 19-25, wherein the site occupancy is the quotient of the measured target abundance and the measured total abundance.

10 [0268] Embodiment 27: The method of any one of embodiments 19-26, wherein the abundance for the NGEP, the internal standard abundance, and the measured abundance of the target glycopeptide are measured in a sample run using mass spectrometry.

[0269] Embodiment 28: A non-transitory computer-readable medium storing computer instructions that, when executed by a computer, cause the computer to perform a method for
15 determining a concentration of a target glycopeptide analyte in a sample, the method comprising: measuring an abundance for a non-glycosylated endogenous peptide (NGEP) of a sample; determining a corrected NGEP abundance for the NGEP using an external standard; measuring an internal standard abundance for an internal standard having a known internal standard concentration; determining a NGEP concentration as a function of the corrected
20 NGEP abundance, the known internal standard concentration, and the internal standard abundance; determining a site occupancy of a target glycopeptide analyte of the sample as a function of a measured target abundance of the target glycopeptide analyte at a given site, and a measured total abundance of all glycopeptides at the same site; and determining a target glycopeptide analyte concentration as a function of the NGEP concentration and the
25 normalized abundance.

[0270] Embodiment 29: The method of embodiment 28, further comprising analyzing the target glycopeptide analyte concentration to generate a treatment for a subject.

[0271] Embodiment 30: The method of embodiments 28 or 29, further comprising analyzing the target glycopeptide analyte concentration to generate a diagnosis for a subject.

30 [0272] Embodiment 31: The method of any one of embodiments 28-30, wherein the NGEP concentration comprises a ratio of the corrected NGEP abundance and the internal standard abundance

[0273] Embodiment 32: The method of embodiment 31, wherein the NGEP concentration is a product of the ratio and the known internal standard concentration.

[0274] Embodiment 33: The method of any one of embodiments 28-32, wherein the site occupancy is determined when three or more glycans are identified at the given site.

5 [0275] Embodiment 34: The method of any one of embodiments 28-33, wherein the target glycopeptide analyte concentration is a product of the NGEP concentration and the normalized abundance.

[0276] Embodiment 35: The method of any one of embodiments 28-34, wherein the site occupancy is the quotient of the measured target abundance and the measured total abundance.

10 [0277] Embodiment 36: The method of any one of embodiments 28-35, wherein the abundance for the NGEP, the internal standard abundance, and the measured abundance of the target glycopeptide are measured in a sample run using mass spectrometry.

[0278] Embodiment 37: A method for preparing samples for a liquid chromatography/mass spectrometry (LC-MS) sample run and normalizing abundance data of a plurality of sample runs, the method comprising: preparing run samples for the plurality of sample runs for LC-MS analysis, the preparing comprising: generating a first set of run samples, each of the first set of run samples including an external standard; and generating a second set of run samples, each of the second set of run samples including at least two target glycopeptide analytes and an internal standard; combining at least two run samples of the first set of run samples to create
15 a pooled standard; analyzing each run sample according to a run order, wherein the run order specifies a relative order of analysis for the run samples of the sample run; and normalizing the raw abundance data of the plurality of sample runs using the pooled standard of each sample run as a reference.

[0279] Embodiment 38: The method of embodiment 37, wherein a first run sample of the first set of run samples occupies an adjacent position in the run order to a second run sample of the second set of run samples.
25

[0280] Embodiment 39: The method of embodiment 37, wherein a run sample of the first set of run samples occupies a position in the run order before all of the run samples of the second set of run samples.

30 [0281] Embodiment 40: The method of embodiment 37, wherein a run sample of the first set of run samples occupies a position in the run order after all of the run samples of the second set of run samples.

[0282] Embodiment 41: The method of embodiment 37, wherein a run sample of the first set of run samples occupies one position for at least every five positions in the run order.

[0283] Embodiment 42: The method of embodiment 37, wherein a run sample of the first set of run samples occupies one position for at least every ten positions in the run order.

5 [0284] Embodiment 43: The method of embodiment 37, wherein the run order comprises a single run sample of the first set of run samples positioned at least every fifteen positions in the run order.

[0285] Embodiment 44: The method of embodiment 37, wherein the run order comprises a single run sample of the first set of run samples followed by a range of between two and
10 twenty-nine run samples of the second set of run samples.

[0286] Embodiment 45: The method of embodiment 37, wherein the run order comprises a single run sample of the first set of run samples followed by a range of between four and fourteen run samples of the second set of run samples.

[0287] Embodiment 46: The method of any one of embodiments 37-45, further comprising:
15 positioning at least one run sample comprising BSA before all the run samples of the first set of run samples and the second set of run samples in the run order.

[0288] Embodiment 47: The method of any one of embodiments 37-46, further comprising: positioning at least one run sample comprising BSA after all the run samples of the first set of run samples and the second set of run samples in the run order.

20 [0289] Embodiment 48: The method of any one of the embodiments 37-47, further comprising: positioning at least one blank run sample before a run sample of the first set of run samples or the second set of run samples in the run order.

[0290] Embodiment 49: The method of any one of the claims 37-48, further comprising: positioning a blank run sample one position before a run sample of the first set of run samples
25 in the run order.

[0291] Embodiment 50: The method of any one of embodiments 37-49, wherein the step of preparing further comprises: enzymatically digesting glycoprotein structures of the run samples to generate the external standard of each run sample of the first set of run samples and the internal standard and the at least two target glycopeptide analytes of each run sample of the
30 second set of run samples.

[0292] Embodiment 51: The method of embodiment 50, further comprising: normalizing raw abundance data generated from each of the at least two target glycopeptide analytes of a same run sample of the second set of run samples using the internal standards.

[0293] Embodiment 52: The method of any one of embodiments 37-51, wherein the external standard comprises a non-glycosylated endogenous peptide (NGEP).

[0294] Embodiment 53: The method of any one of embodiments 37-52, wherein each of the at least two target glycopeptide analytes has a corresponding internal standard for a reference.

[0295] Embodiment 54: The method of any one of embodiments 37-53, wherein the internal standard of at least one run sample of the second set of run samples is a surrogate.

[0296] Embodiment 55: The method of any one of embodiments 37-54, wherein a source of the external standard includes serum.

[0297] Embodiment 56: The method of any one of embodiments 37-55, wherein an external standard of the first set of run samples includes at least one pooled plasma digest when a run sample in the second set of runs samples includes a patient plasma sample.

[0298] Embodiment 57: The method of embodiment 56, wherein the patient plasma sample includes a target analyte.

[0299] Embodiment 58: The method of any one of embodiments 37-57, wherein the concentration of the external standard is unknown prior to analyzing.

[0300] Embodiment 59: The method of any one of embodiments 37-58, wherein the concentration of the pooled standard is unknown prior to analyzing.

[0301] Embodiment 60: A sample processing system for carrying out a process for preparing samples for a liquid chromatography/mass spectrometry (LC-MS) sample run and normalizing abundance data of a plurality of sample runs, the sample processing system comprising: a sample preparation system, including a fluidic instrument, for performing a process comprising: generating a first set of run samples, each of the first set including an external standard; generating a second set of run samples, each of the second set including at least two target glycopeptide analytes and an internal standard; and combining at least two run samples of the first set to create a pooled standard; a sample analysis system, including a LC-MS instrument, for analyzing glycopeptide analytes, the process comprising analyzing each run sample according to a run order, wherein the run order is stored on a data store in electronic communication with the LC-MS instrument and specifies a relative order of analysis for the run samples of the sample run; and a data analysis system, including a normalization module of a peptide structure analyzer, for normalizing raw abundance data of the plurality of sample runs using the pooled standard of each sample run as a common reference.

[0302] Embodiment 61: The system of embodiment 60, wherein a first run sample of the first set of run samples occupies an adjacent position in the run order to a second run sample of the second set of run samples.

5 [0303] Embodiment 62: The system of embodiment 60, wherein a run sample of the first set of run samples occupies a position before all of the run samples of the second set of run samples in the run order.

[0304] Embodiment 63: The system of embodiment 60, wherein a run sample of the first set of run samples occupies a position after all of the run samples of the second set of run samples in the run order.

10 [0305] Embodiment 64: The system of embodiment 60, wherein the run order comprises a single run sample of the first set of run samples positioned at least every five positions in the run order.

[0306] Embodiment 65: The system of embodiment 60, wherein the run order comprises a single run sample of the first set of run samples positioned at least every ten positions in the run order.

[0307] Embodiment 66: The system of claim 60, wherein the run order comprises a single run sample of the first set of run samples positioned at least every fifteen positions in the run order.

20 [0308] Embodiment 67: The system of embodiment 60, wherein the run order comprises a single run sample of the first set of run samples followed by a range of between two and twenty-nine run samples of the second set.

[0309] Embodiment 68: The system of embodiment 60, wherein the run order comprises a single run sample of the first set of run samples followed by a range of between four and fourteen run samples of the second set of run samples.

25 [0310] Embodiment 69: The system of any one of embodiments 60-68, wherein at least one run sample comprising BSA is positioned before all the run samples of the first set of run samples and the second set of run samples in the run order.

[0311] Embodiment 70: The system of any one of embodiments 60-69, wherein at least one run sample comprising BSA is positioned after all the run samples of the first set of run samples and the second set of run samples in the run order.

30 [0312] Embodiment 71: The system of any one of embodiments 60-70, wherein at least one blank run sample is positioned before a run sample of the first set of run samples or the second set of run samples in the run order.

[0313] Embodiment 72: The system of any one of embodiments 60-71, wherein a blank run sample is positioned one position before a run sample of the first set of run samples in the run order.

5 [0314] Embodiment 73: The system of any one of embodiments 60-72, wherein the process performed using the sample preparation system further comprises enzymatically digesting glycoprotein structures of the run samples to generate the external standard of each run sample of the first set of run samples and the internal standard and the at least two target glycopeptide analytes of each run sample of the second set of run.

10 [0315] Embodiment 74: The system of any one of embodiments 60-73, wherein the data analysis system is configured to normalize the raw abundance data generated from each of the at least two target glycopeptide analytes of a same run sample using the internal standard.

[0316] Embodiment 75: The system of any one of embodiments 60-74, wherein the external standard comprises a non-glycosylated endogenous peptide (NGEP).

15 [0317] Embodiment 76: The system of any one of embodiments 60-75, wherein the external standard comprises a non-glycosylated endogenous peptide (NGEP).

[0318] Embodiment 77: The system of any one of embodiments 60-76, wherein each of the at least two target glycopeptide analytes has a corresponding internal standard for a reference.

20 [0319] Embodiment 78: The system of any one of embodiments 60-77, wherein the internal standard of at least one run sample of the second set of run samples is a surrogate.

[0320] Embodiment 79: The system of any one of embodiments 60-78, wherein a source of the external standard includes serum.

25 [0321] Embodiment 80: The system of any one of embodiments 60-79, wherein an external standard of the first set of run samples includes at least one pooled plasma digest when a run sample in the second set of runs samples includes a patient plasma sample.

[0322] Embodiment 81: The system of embodiment 80, wherein the patient plasma sample includes a target analyte.

[0323] Embodiment 82: The system of any one of embodiments 60-81, wherein the concentration of the external standard is unknown prior to analyzing.

30 [0324] Embodiment 83: The system of any one of embodiments 60-82, wherein the concentration of the pooled standard is unknown prior to analyzing.

[0325] Embodiment 84: The method of any one of embodiments 1-9 and 19-27, further comprising a method for preparing samples for a LC-MS sample run and normalizing abundance data of a plurality of sample runs, of any one of embodiments 37-59.

[0326] Embodiment 85: The method of any one of embodiments 10-18 and 28-36, further
5 comprising a method for preparing samples for a LC-MS sample run and normalizing abundance data of a plurality of sample runs, of any one of embodiments 37-59.

CLAIMS

What is claimed is:

1. A method of determining a concentration of a target glycopeptide analyte in a sample, the
5 method comprising:
 - measuring an abundance for a non-glycosylated endogenous peptide (NGEP) of a sample;
 - determining a corrected NGEP abundance for the NGEP using an external standard;
 - measuring an internal standard abundance for an internal standard having a known
10 internal standard concentration;
 - determining an NGEP concentration for the NGEP as a function of the corrected NGEP abundance, the known internal standard concentration, and the internal standard abundance;
 - determining a normalized abundance of a target glycopeptide analyte of the sample as
15 a function of a measured abundance of the target glycopeptide analyte and the measured abundance of the NGEP on the same protein as the target glycoprotein analyte; and
 - determining a target glycopeptide analyte concentration as a function of the NGEP concentration and the normalized abundance.
2. The method of claim 1, further comprising analyzing the target glycopeptide analyte
20 concentration to generate a treatment for a subject.
3. The method of claims 1 or 2, further comprising analyzing the target glycopeptide analyte concentration to generate a diagnosis for a subject.
4. The method of any one of claims 1-3, wherein the NGEP concentration comprises a ratio of the corrected NGEP abundance and the internal standard abundance.
- 25 5. The method of claim 4, wherein the NGEP concentration is a product of the ratio and the known internal standard concentration.
6. The method of any one of claims 1-5, wherein the normalized abundance is determined when one or two glycans are identified at a target glycosylation site on the protein.
7. The method of any one of claims 1-6, wherein the target glycopeptide analyte
30 concentration is a product of the NGEP concentration and the normalized abundance.

8. The method of any one of claims 1-7, wherein the normalized abundance is the quotient of the first measured abundance and the second measured abundance.
9. The method of any one of claims 1-8, wherein the abundance for the NGEP, the internal standard abundance, and the measured abundance of the target glycopeptide are measured
5 in a sample run using mass spectrometry.
10. A method for preparing samples for a liquid chromatography/mass spectrometry (LC-MS) sample run and normalizing abundance data of a plurality of sample runs, the method comprising:
preparing run samples for the plurality of sample runs for LC-MS analysis, the
10 preparing comprising:
generating a first set of run samples, each of the first set of run samples
including an external standard; and
generating a second set of run samples, each of the second set of run samples
including at least two target glycopeptide analytes and an internal standard;
15 combining at least two run samples of the first set of run samples to create a pooled standard;
analyzing each run sample according to a run order, wherein the run order specifies a relative order of analysis for the run samples of the sample run; and
normalizing the raw abundance data of the plurality of sample runs using the pooled
20 standard of each sample run as a reference.
11. The method of claim 10, wherein a first run sample of the first set of run samples occupies an adjacent position in the run order to a second run sample of the second set of run samples.
12. The method of claim 10, wherein a run sample of the first set of run samples occupies a
25 position in the run order before all of the run samples of the second set of run samples.
13. The method of claim 10, wherein a run sample of the first set of run samples occupies a position in the run order after all of the run samples of the second set of run samples.
14. The method of claim 10, wherein a run sample of the first set of run samples occupies one position for at least every five positions in the run order.

15. The method of claim 10, wherein a run sample of the first set of run samples occupies one position for at least every ten positions in the run order.
16. The method of claim 10, wherein the run order comprises a single run sample of the first set of run samples positioned at least every fifteen positions in the run order.
- 5 17. The method of claim 10, wherein the run order comprises a single run sample of the first set of run samples followed by a range of between two and twenty-nine run samples of the second set of run samples.
18. The method of claim 10, wherein the run order comprises a single run sample of the first set of run samples followed by a range of between four and fourteen run samples of the
10 second set of run samples.
19. The method of any one of claims 10-18, further comprising:
positioning at least one run sample comprising BSA before all the run samples of the first set of run samples and the second set of run samples in the run order.
20. The method of any one of claims 10-19, further comprising:
15 positioning at least one run sample comprising BSA after all the run samples of the first set of run samples and the second set of run samples in the run order.
21. The method of any one of the claims 10-20, further comprising:
positioning at least one blank run sample before a run sample of the first set of run samples or the second set of run samples in the run order.
- 20 22. The method of any one of the claims 10-21, further comprising:
positioning a blank run sample one position before a run sample of the first set of run samples in the run order.
23. The method of any one of claims 10-22, wherein the step of preparing further comprises:
enzymatically digesting glycoprotein structures of the run samples to generate the
25 external standard of each run sample of the first set of run samples and the internal standard and the at least two target glycopeptide analytes of each run sample of the second set of run samples.
24. The method of claim 23, further comprising:

normalizing raw abundance data generated from each of the at least two target glycopeptide analytes of a same run sample of the second set of run samples using the internal standards.

25. The method of any one of claims 10-24, wherein the external standard comprises a non-glycosylated endogenous peptide (NGEP).
26. The method of any one of claims 10-25, wherein each of the at least two target glycopeptide analytes has a corresponding internal standard for a reference.
27. The method of any one of claims 10-26, wherein the internal standard of at least one run sample of the second set of run samples is a surrogate.
28. The method of any one of claims 10-27, wherein a source of the external standard includes serum.
29. The method of any one of claims 10-28, wherein an external standard of the first set of run samples includes at least one pooled plasma digest when a run sample in the second set of runs samples includes a patient plasma sample.
30. The method of claim 29, wherein the patient plasma sample includes a target analyte.
31. The method of any one of claims 10-30, wherein the concentration of the external standard is unknown prior to analyzing.
32. The method of any one of claims 10-31, wherein the concentration of the pooled standard is unknown prior to analyzing.
33. A sample processing system for carrying out a process for preparing samples for a liquid chromatography/mass spectrometry (LC-MS) sample run and normalizing abundance data of a plurality of sample runs, the sample processing system comprising:
a sample preparation system, including a fluidic instrument, for performing a process comprising:
generating a first set of run samples, each of the first set including an external standard;
generating a second set of run samples, each of the second set including at least two target glycopeptide analytes and an internal standard; and
combining at least two run samples of the first set to create a pooled standard;

a sample analysis system, including a LC-MS instrument, for analyzing glycopeptide analytes, the process comprising analyzing each run sample according to a run order, wherein the run order is stored on a data store in electronic communication with the LC-MS instrument and specifies a relative order of analysis for the run samples of the sample run; and

a data analysis system, including a normalization module of a peptide structure analyzer, for normalizing raw abundance data of the plurality of sample runs using the pooled standard of each sample run as a common reference.

34. The system of claim 33, wherein a first run sample of the first set of run samples occupies an adjacent position in the run order to a second run sample of the second set of run samples.
35. The system of claim 33, wherein a run sample of the first set of run samples occupies a position before all of the run samples of the second set of run samples in the run order.
36. The system of claim 33, wherein a run sample of the first set of run samples occupies a position after all of the run samples of the second set of run samples in the run order.
37. The system of claim 33, wherein the run order comprises a single run sample of the first set of run samples positioned at least every five positions in the run order.
38. The system of claim 33, wherein the run order comprises a single run sample of the first set of run samples positioned at least every ten positions in the run order.
39. The system of claim 33, wherein the run order comprises a single run sample of the first set of run samples positioned at least every fifteen positions in the run order.
40. The system of claim 33, wherein the run order comprises a single run sample of the first set of run samples followed by a range of between two and twenty-nine run samples of the second set.
41. The system of claim 33, wherein the run order comprises a single run sample of the first set of run samples followed by a range of between four and fourteen run samples of the second set of run samples.
42. The system of any one of claims 33-41, wherein at least one run sample comprising BSA is positioned before all the run samples of the first set of run samples and the second set of run samples in the run order.

43. The system of any one of claims 33-42, wherein at least one run sample comprising BSA is positioned after all the run samples of the first set of run samples and the second set of run samples in the run order.
44. The system of any one of claims 33-43, wherein at least one blank run sample is
5 positioned before a run sample of the first set of run samples or the second set of run samples in the run order.
45. The system of any one of claims 33-44, wherein a blank run sample is positioned one position before a run sample of the first set of run samples in the run order.
46. The system of any one of claims 33-45, wherein the process performed using the sample
10 preparation system further comprises enzymatically digesting glycoprotein structures of the run samples to generate the external standard of each run sample of the first set of run samples and the internal standard and the at least two target glycopeptide analytes of each run sample of the second set of run.
47. The system of any one of claims 33-46, wherein the data analysis system is configured to
15 normalize the raw abundance data generated from each of the at least two target glycopeptide analytes of a same run sample using the internal standard.
48. The system of any one of claims 33-47, wherein the external standard comprises a non-glycosylated endogenous peptide (NGEP).
49. The system of any one of claims 33-48, wherein each of the at least two target
20 glycopeptide analytes has a corresponding internal standard for a reference.
50. The system of any one of claims 33-49, wherein the internal standard of at least one run sample of the second set of run samples is a surrogate.
51. The system of any one of claims 33-50, wherein a source of the external standard includes serum.
52. The system of any one of claims 33-51, wherein an external standard of the first set of run
25 samples includes at least one pooled plasma digest when a run sample in the second set of runs samples includes a patient plasma sample.
53. The system of claim 52, wherein the patient plasma sample includes a target analyte.

54. The system of any one of claims 33-53, wherein the concentration of the external standard is unknown prior to analyzing.
55. The system of any one of claims 33-54, wherein the concentration of the pooled standard is unknown prior to analyzing.
- 5 56. The method of any one of claims 1-9, further comprising a method for preparing samples for a LC-MS sample run and normalizing abundance data of a plurality of sample runs, of any one of claims 10-32.