(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 August 2023 (10.08.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/150729 A2

(51) International Patent Classification:
G16B 20/00 (2019.01)

(21) International Application Number:
PCT/US2023/062001

(22) International Filing Date:
03 February 2023 (03.02.2023)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

63/306,426	03 February 2022 (03.02.2022)	US
63/269,804	23 March 2022 (23.03.2022)	US
63/269,884	24 March 2022 (24.03.2022)	US
63/362,303	31 March 2022 (31.03.2022)	US
63/365,850	03 June 2022 (03.06.2022)	US
63/479,909	13 January 2023 (13.01.2023)	US

(71) Applicant: **VENN BIOSCIENCES CORPORATION** [US/US]; 2 Tower Place, 5th Floor, South San Francisco, California 94080 (US).

(72) Inventors: **CHEN, Shao-Yung**; 518 Palm Avenue, South San Francisco, California 94080 (US). **LIN, Chi-Hung**; 2302 Carlmont Dr. Apt 5, Belmont, California 94002 (US). **GANDHI, Monil Dilip**; c/o VENN BIOSCIENCES CORPORATION, 2 Tower Place, 5th Floor, South San Francisco, California 94080 (US). **SHIPMAN, III, Richard David**; c/o VENN BIOSCIENCES CORPORATION, 2 Tower Place, 5th Floor, South San Francisco, California 94080 (US). **WALTON, Kelly F.**; c/o VENN BIOSCIENCES CORPORATION, 2 Tower Place, 5th Floor, South San Francisco, California 94080 (US).

(54) Title: SYSTEMS AND METHODS FOR GLYCOMOLECULE IDENTIFICATION, SEARCH AND COMPARING AND ANALYZING THE RESULTS THEREOF

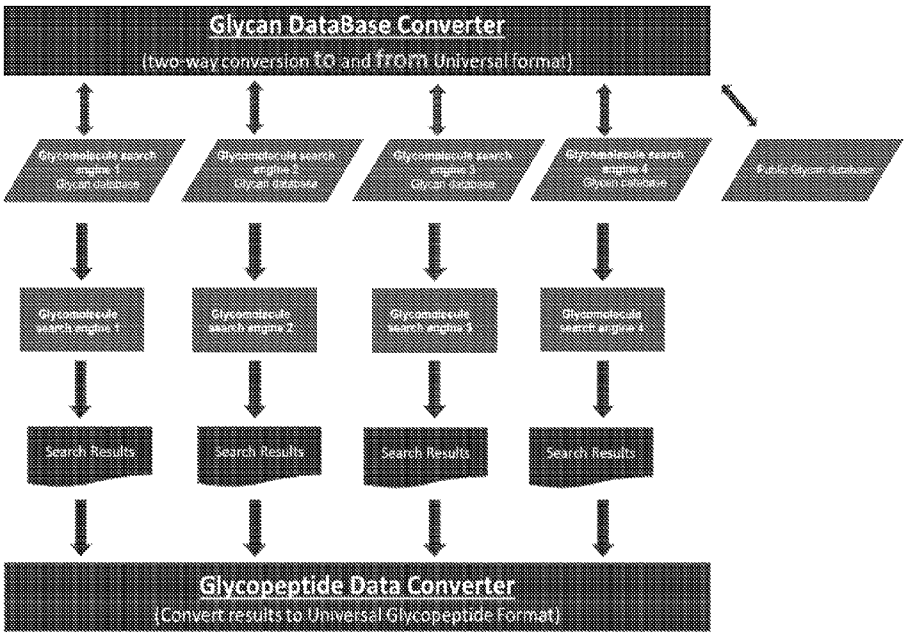


FIG. 1

(57) Abstract: Set forth herein is a system and method for converting glycan representations between different platform-specific glycan formats via a universal glycan format. The use of a universal glycan format for conversion between different platform-specific glycan formats, such as different glycomolecule search engine-specific formats, can reduce the number of conversion rule sets required to convert glycan data formats among the different platform-specific glycan formats. The universal glycan format can also improve readability, and simplify analysis of glycomolecules. Also provided is a system and method for comparing or analyzing glycomolecule search results from different glycomolecule search engines and a user interface, system and method for curating a consensus list of glycomolecules (e.g., glycopeptides, glycoDNA, glycoRNA, glycolipids) identified in a biological sample on a user interface from glycomolecule search results sets from multiple glycomolecule search engines, where the glycomolecule search results sets may include

WO 2023/150729 A2

[Continued on next page]

(74) **Agent: LOZAN, Vladimir, S.**; Knobbe, Martens, Olson & Bear, LLP, 2040 Main Street, 14th Floor, Irvine, CA 92614 (US).

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

conflicting identifications between the glycomolecule search engines.

SYSTEMS AND METHODS FOR GLYCOMOLECULE IDENTIFICATION, SEARCH AND COMPARING AND ANALYZING THE RESULTS THEREOF

BACKGROUND

Field

[0001] The instant disclosure is directed to a system for managing, searching, curating, comparing and/or analyzing sample data and/or search results associated with glycans, glycopeptides, or other glycomolecules (e.g., glycopeptides, glycoDNA, glycoRNA, glycolipids) that are predicted to be present in a sample.

Description of the Related Art

[0002] Studies involving glycans, glycopeptides, or other glycomolecules in samples often require the use of search engines to quantify and/or identify such glycans, glycopeptides, or other glycomolecules present in the samples. For purposes of this disclosure, "glycomolecules" include glycans and glycoconjugates (molecules that are conjugated to glycans, e.g., glycopeptides, glycoproteins, glycolipids, glycoRNA, glycoDNA). There are currently a number of such search engines. For example, search engines under the names BYONIC, PGLYCO3, MSFRAGGER-GLYCO, METAMORPHEUS (also known as O-PAIR), STRUCGP, and other similar search engines can be used to match a mass spectrometry signature to known glycomolecule signatures. The identification process used by these search engines may involve inputting mass spectrometry data, a glycan database, and/or a conjugate database to perform a search. For example, in the case of glycopeptide search process, a researcher may input, into a glycopeptide search engine, a mass spectrometry data file (e.g., an mzML file), a glycan database file, and a protein/peptide database file.

[0003] In many cases, researchers may benefit from leveraging multiple glycan databases to quantify/identify glycomolecules in a sample. One reason for using multiple glycan databases is that different glycan databases contain different subsets of glycan data, and it may be desirable to search among these different subsets. Another reason is that different search engines use different search algorithms and thus yield different results. Thus, it may be desirable to run searches on multiple search engines to validate search results or to gain a deeper understanding of the sample. However, these search engines are not compatible with every glycan database; in fact, they typically have a platform-specific

format that specifies how glycans are to be represented. Thus, only databases that include glycan representations in a platform-specific format corresponding to a particular search engine would be compatible to that search engine. This restricts the scope of the "inputs" (e.g., glycan data from glycan databases) that can be fed into the search engines, and therefore greatly limits what a researcher can do. For example, glycan descriptions in a glycan database using the PGLYCO3 format are different from, and therefore incompatible with, glycan descriptions in a glycan database using the METAMORPHEUS format. Moreover, the formats of the "outputs" of the different search engines may be different. For example, in the case of glycopeptide search engines, the same search conducted on two different glycopeptide search engines may output search results in two different formats, each specific to the respective platform. Many research applications can involve running searches using different glycan databases (the inputs) and/or comparing glycomolecule search results (the outputs) from different search engines. However, given the different input and output formats, this comparison can be exceedingly difficult and time consuming. For example, the same set of raw input data cannot be searched across multiple search engines. Similarly, the outputs of different search engines cannot be directly compared.

[0004] Additionally, since each search engine has a different "confidence score" to rank the confidence and credibility of each result that is returned from a glycomolecule search, it is difficult to compare among the search results from these different search engines. Moreover, different search engines rank their identifications/annotations using different strategies, such that even the same set of raw data can have different glycopeptide annotations based on the search engines used, and these annotations can even be mutually exclusive. This leads to challenges when attempting to search multiple glycan databases and compare the results across the different databases.

SUMMARY

[0005] The systems, devices, kits, and methods disclosed herein each have several aspects, no single one of which is solely responsible for their desirable attributes. Without limiting the scope of the claims, some prominent features will now be discussed briefly. Numerous other embodiments are also contemplated, including embodiments that have fewer, additional, and/or different components, steps, features, objects, benefits, and advantages. The components, aspects, and steps may also be arranged and ordered differently. After considering this discussion, and particularly after reading the section

entitled "Detailed Description," one will understand how the features of the devices and methods disclosed herein provide advantages over other known devices and methods.

[0006] Provided herein is a glycan data management system for handling data associated with glycans, the system comprising: a non-transitory memory comprising instructions for converting among a plurality of platform-specific glycan formats and a universal glycan format by applying at least one of a plurality of format conversion rule sets, the plurality of format conversion rule sets comprising: a first rule set configured to convert glycan representations of a first platform-specific glycan format to and from the universal glycan format; and a second rule set configured to convert glycan representations of a second platform-specific glycan format to and from the universal glycan format; and a processor configured to: access an input glycan data comprising one or more glycan representations in the first platform-specific glycan format, the second platform-specific glycan format, or the universal glycan format; and execute the instructions to convert the input glycan data between: the first platform-specific glycan format and the universal glycan format; the second platform-specific glycan format and the universal glycan format; or the first platform-specific glycan format and the second platform-specific glycan format via the universal glycan format. Optionally, the processor is configured to execute the instructions to convert the input glycan data from the first platform-specific glycan format to the second platform-specific glycan format by applying: (1) the first rule set to convert the input glycan data from the first platform-specific glycan format to the universal glycan format; and (2) the second rule set to convert the input glycan data from the universal glycan format to the second platform-specific glycan format. In some embodiments, the non-transitory memory further comprises instructions for obtaining the input glycan data from a first glycan database in the first platform-specific glycan format, and wherein execution of the instructions by the processor generates a second glycan database comprising glycan data in the second platform-specific glycan format.

[0007] In some embodiments, the first platform-specific glycan format is compatible with a first glycomolecule search engine. Optionally, the glycomolecule search engine comprises a glycopeptide search engine.

[0008] In some embodiments, the second platform-specific glycan format, but not the first platform-specific glycan format, is compatible with a second glycomolecule search engine. Optionally, the universal glycan format is not compatible with the first or the second glycomolecule search engine.

[0009] In some embodiments, the universal glycan format comprises compositional information only. In some embodiments, the universal glycan format comprises a single-letter representation for different monosaccharides. In some embodiments, the universal glycan format comprises a zero for a monosaccharide that is not present in the glycan data. In some embodiments, the universal glycan format comprises: a first number *l* corresponding to an amount of hexose in the glycan data; a second number *m* corresponding to an amount of N-acetylhexosamine in the glycan data; a third number *n* corresponding to an amount of fucose in the glycan data; and a fourth number *o* corresponding to an amount of N-acetylneuraminic acid in the glycan data, wherein *l*, *m*, *n*, and *o* are each an integer 0 or greater. Optionally, the universal glycan format further comprises a fifth number *p* corresponding to an amount of N-glycolylneuraminic acid (Neu5Gc), wherein *p* is an integer 0 or greater.

[0010] In some embodiments, at least one of the first and second platform-specific glycan formats comprises glycan structural information.

[0011] Also provided is a method of expanding the cross-compatibility of a glycan database, comprising: providing to the glycan data management system of any one of claims 1-11 a glycan database comprising input glycan data comprising glycan representations in a first platform-specific glycan format; and adding to the plurality of format conversion rule sets a third rule set configured to convert glycan representations of the third platform-specific glycan format to and from the universal glycan format, wherein none of the other rule sets among the plurality of format conversion rule sets is configured to convert glycan representations of the third platform-specific glycan format to and from the universal glycan format, wherein the third platform-specific glycan format is compatible with a third glycomolecule search engine and wherein the first platform-specific glycan format is not compatible with the third glycomolecule search engine. Optionally, the method further includes causing the processor to access the input glycan data of the glycan database and executing the instructions to convert the input glycan data to the third platform-specific glycan format via the universal glycan format.

[0012] Provided herein is a method of converting a glycan data format, comprising: receiving first glycan data comprising one or more glycan representations in a first platform-specific glycan format; and converting the first glycan data from the first platform-specific glycan format to a second platform-specific glycan format via a universal glycan format, wherein the second platform-specific glycan format, but not the first platform-specific glycan format, is compatible with a first glycomolecule search engine.

[0013] In some embodiments, the first glycomolecule search engine comprises a first glycopeptide search engine. In some embodiments, the first platform-specific glycan format is compatible with a second glycomolecule search engine. In some embodiments, the second glycomolecule search engine comprises a second glycopeptide search engine.

[0014] In some embodiments, converting the first glycan data comprises: determining that the first glycan data comprises the one or more glycan representations in the first platform-specific glycan format; and applying to the first glycan data: (1) a first rule set for converting the first platform-specific glycan format to the universal glycan format; and (2) a second rule set for converting the universal glycan data format to the second platform-specific glycan format.

[0015] In some embodiments, the universal glycan format comprises compositional information only. In some embodiments, the universal glycan format comprises a single-letter representation for different monosaccharides. In some embodiments, the universal glycan format comprises a zero for a monosaccharide that is not present in the glycan data. In some embodiments, the universal glycan format comprises: a first number *l* corresponding to an amount of hexose in the glycan data; a second number *m* corresponding to an amount of N-acetylhexosamine in the glycan data; a third number *n* corresponding to an amount of fucose in the glycan data; and a fourth number *o* corresponding to an amount of N-acetylneuraminic acid in the glycan data, wherein *l*, *m*, *n*, and *o* are each an integer 0 or greater. Optionally, the universal glycan data format further comprises a fifth number *p* corresponding to an amount of N-glycolylneuraminic acid, wherein *p* is an integer 0 or greater.

[0016] In some embodiments, at least one of the first and second platform-specific glycan formats comprises structural information.

[0017] Also provided herein is a glycan data management system for handling data associated with glycans, the system comprising: a processor; and non-transitory memory comprising instructions, which when executed by the processor causes the processor to perform any of the method of the present disclosure.

[0018] Also provided is a method of converting a glycan format, comprising: receiving glycan data comprising a glycan representation in a first format, the first format comprising compositional information; and converting the glycan representation to a second format, the second format comprising compositional information, wherein either: (1) the first format further comprises glycan structural information, and the second format comprises compositional information only, and wherein converting the glycan

representation comprises enumerating an amount of each of the monosaccharides in the glycan; or (2) the first format comprises compositional information only, and the second format further comprises glycan structural information, and wherein converting the representation comprises: receiving information about a source of the glycan data; and generating a list of potential glycan structures based on the compositional information and the source of the glycan data. In some embodiments, the source comprises human urine, human stool, human blood, or human serum. In some embodiments, the source comprises a yeast, a plant, a non-human mammal, or a human.

[0006] Provided herein is a glycomolecule data analysis system for comparing glycomolecule search outputs from two or more glycomolecule search engines, the system comprising: a processor; and a non-transitory memory configured to cause the processor to: access a first search output associated with a first glycomolecule search engine and a second search output associated with a second glycomolecule search engine, wherein the first search output comprises first glycomolecular data for each of one or more identified first glycomolecules in a first format, and wherein the second search output comprises second glycomolecular data for each of one or more identified second glycomolecules in a second format; convert one or more of the first and second glycomolecular data to a common format, wherein the common format comprises the first format, the second format, or a universal glycomolecular data format; and perform a comparison between the first and second glycomolecular data in the common format. Optionally, one or more of the first and second glycomolecular data is converted to the common format by: extracting one or more parameters associated with each of the one or more identified first and second glycomolecules in the first and second glycomolecular data, respectively, wherein the one or more parameters are in a search engine-specific format and comprises one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score; and converting the one or more parameters associated with each of the one or more identified first and second glycomolecules to the common format. In some embodiments, the common format is the universal glycomolecular data format.

[0007] In some embodiments, the first and second search outputs correspond to the same glycomolecule mass spectrometry data run on the first and second glycomolecule search engines. In some embodiments, the non-transitory memory is configured to cause the processor to: access a third search output associated with a third glycomolecule search engine, wherein the third search output comprises third glycomolecular data for each of

one or more identified third glycomolecules in a third format; convert one or more of the first, second, and third glycomolecular data to a common format, wherein the common format comprises the first format, the second format, the third format, or the universal glycomolecular data format; and perform a comparison between any two or more of the first, second, and third glycomolecular data in the common format. Optionally, the first, second, and third search outputs correspond to the same glycomolecule mass spectrometry data run on the first, second, and third glycomolecule search engines, respectively. In some embodiments, the third glycomolecular data is converted to the universal glycomolecular data format by extracting one or more parameters associated with each of the one or more identified third glycomolecules in the third glycomolecular data, wherein the one or more parameters comprises one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score; and converting the one or more parameters associated with each of the one or more identified third glycomolecules to the common format.

[0008] Also provided is a glycomolecule data analysis system for comparing glycomolecule search outputs from two or more glycomolecule search engines, the system comprising: a processor; and a non-transitory memory configured to cause the processor to: access two or more search outputs from two or more glycomolecule search engines, each search output (1) being associated with one of the two or more glycomolecule search engines and (2) comprising glycomolecular data in a search engine-specific format for each of one or more glycomolecules identified by the glycomolecule search engine; extract one or more parameters associated with each of the one or more identified glycomolecules in the glycomolecular data of each of the two or more search outputs, wherein the one or more parameters comprises an identifier; and perform a comparison between the glycomolecular data from the two or more search outputs based on the one or more extracted parameters. Optionally, the comparison is performed by identifying glycomolecules associated with the same identifier in the glycomolecular data from the two or more search outputs. In some embodiments, the one or more parameters further comprises one or more of: glycan data, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score.

[0009] In some embodiments, the two or more search outputs correspond to the same glycomolecule mass spectrometry data run on the two or more glycomolecule search engines, respectively. In some embodiments, the non-transitory memory is further configured to cause the processor to convert glycomolecular data of one or more of the two

or more search outputs to a common format, wherein the common format comprises one of the search engine-specific formats or a universal glycomolecular data format.

[0010] In some embodiments, the common format is the universal glycomolecular data format. In some embodiments, the universal glycomolecular data format comprises identifying information for the glycomolecule search engine. In some embodiments, the universal glycomolecular data format comprises, for each identified glycomolecule, one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score.

[0011] In some embodiments, the comparison comprises identifying an overlap between any two or more of the glycomolecular data. Optionally, the overlap comprises a set of glycomolecules that are identified in any two or more of the glycomolecular data. In some embodiments, the overlap comprises one or more of peptide sequences, protein IDs, retention times, m/z values, or charge numbers.

[0012] In some embodiments, the comparison comprises identifying a difference between any two or more of the glycomolecular data. Optionally, the difference comprises one or more sets of glycomolecules that are identified in one of the any two or more of the glycomolecular data as being different from the other one of the any two or more of the glycomolecular data. In some embodiments, the difference comprises one or more of peptide sequences, protein IDs, retention times, m/z values, or charge numbers.

[0013] In some embodiments, the glycomolecule is a glycopeptide, glycoprotein, glycolipid, glycoRNA, or glycoDNA. In some embodiments, the glycomolecule search engine is a glycopeptide search engine, glycoprotein search engine, glycolipid search engine, glycoDNA search engine, glycoRNA search engine, or glycoDNA search engine.

[0014] Also provided herein is a method of comparing glycomolecule search outputs from two or more glycomolecule search engines, comprising: receiving a first search output associated with a first glycomolecule search engine and a second search output associated with a second glycomolecule search engine, wherein the first search output comprises first glycomolecular data for each of one or more identified first glycomolecules in a first format, and wherein the second search output comprises second glycomolecular data for each of one or more identified second glycomolecules in a second format; converting one or more of the first and second glycomolecular data to a common format, wherein the common format comprises the first format, the second format, or a universal glycomolecular data format; and comparing the first and second glycomolecular data in the common format. Optionally, the first and second glycomolecular data comprise

one or more parameters associated with each of the one or more identified first and second glycomolecules in the first and second glycomolecular data, respectively, wherein the one or more parameters are in a search engine-specific format and comprises one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score, and wherein converting the one or more of the first and second glycomolecular data to the common format comprises converting the one or more parameters to the common format. In some embodiments, the first and second search outputs correspond to the same glycomolecule mass spectrometry data run on the first and second glycomolecule search engines. Optionally, the method further comprises providing a data set for running on each of the two or more glycomolecule search engines, wherein the data sets comprise the same glycomolecule mass spectrometry data. In some embodiments, the method further comprises: generating the glycomolecule mass spectrometry data from a biological sample; and providing the data set comprising the generated glycomolecule mass spectrometry data. Optionally, the biological sample is from human urine, human stool, human blood, or human serum.

[0015] In some embodiments, the two or more glycomolecule search engines are two or more glycopeptide search engines, the method further comprising providing a data set for running on each of the two or more glycopeptide search engines, wherein the data sets comprise the same glycopeptide mass spectrometry data. Optionally, each data set comprises a glycan database comprising glycan data in a glycan format that is compatible with a corresponding glycopeptide search engine of the two or more glycopeptide search engines. Optionally, providing the data set further comprises: providing the glycan database in a universal glycan format; and converting the glycan database to a glycan format compatible with one of the two or more glycopeptide search engines. Optionally, providing the data set further comprises: receiving the glycan database in a glycan format compatible with one of the two or more glycopeptide search engines; and converting the glycan database to the universal glycan format.

[0016] In some embodiments, the method further comprises: receiving a third search output associated with a third glycomolecule search engine, wherein the third search output comprises third glycomolecular data for each of one or more identified third glycomolecules in a third format; converting one or more of the first, second, and third glycomolecular data to a common format, wherein the common format comprises the first format, the second format, the third format, or the universal glycomolecular data format; and comparing any two or more of the first, second, and third glycomolecular data, each in

the common format. Optionally, the first, second, and third search outputs correspond to the same glycomolecule mass spectrometry data run on the first, second, and third glycomolecule search engines.

[0017] In some embodiments, the common format is the universal glycomolecular data format. In some embodiments, the universal glycomolecular data format comprises identifying information for the glycomolecule search engine. In some embodiments, the universal glycopeptide data format comprises, for each identified glycomolecule, one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, or a confidence score. In some embodiments, the glycomolecule is a glycopeptide, glycoprotein, glycolipid, glycoRNA, or glycoDNA. In some embodiments, the glycomolecule search engine is a glycopeptide search engine, glycoprotein search engine, glycolipid search engine, glycoRNA search engine, or glycoDNA search engine.

[0018] Also provided is a method for determining a glycomolecule profile of a biological sample, comprising: providing to each of two or more glycomolecule search engines a data set comprising glycomolecule mass spectrometry data obtained from a biological sample, wherein the same glycomolecule mass spectrometry data are provided to each of the two or more glycomolecule search engines; receiving an output from each of the two or more glycomolecule search engines run on the provided data set, wherein each output from the two or more glycomolecule search engines comprises glycomolecular data for each of one or more identified glycomolecules in a search engine-specific format; converting glycomolecular data of one or more of the outputs from the two or more glycomolecule search engines to a common format, wherein the common format comprises one of the search engine-specific formats or a universal glycomolecular data format; and determining a glycomolecule profile of the sample based on the glycomolecular data of the outputs from the two or more glycomolecule search engines in the common format.

[0019] Also provided is an electronic system for comparing glycomolecule search outputs from two or more glycomolecule search engines, comprising: a processor; and non-transitory memory comprising instructions, which when executed by the processor causes the processor to perform a method of the present disclosure.

[0006] Provided herein is a method of identifying glycomolecules in a biological sample, comprising: receiving a first search output for a biological sample, the first search output being associated with a first glycomolecule search engine; determining that the first search output identifies a first set of glycomolecules as being of a first type

and a second set of glycomolecules as being of a second type; receiving a second search output for the biological sample, the second search output being associated with a second glycomolecule search engine; determining that the second search output identifies a third set of glycomolecules as being of the first type and a fourth set of glycomolecules as being of the second type; and identifying among the first and second search outputs a conflicting subset of glycomolecules, wherein the conflicting subset of glycomolecules comprises: (a) glycomolecules that are present in both the first set and the fourth set; or (b) glycomolecules that are present in both the second set and the third set. Optionally, the first set of glycomolecules is mutually exclusive of the third set of glycomolecules and the second set of glycomolecules is mutually exclusive of the fourth set of glycomolecules. In some embodiments, the first search output comprises: a first set of identifiers associated with each glycomolecule of the first set of glycomolecules; and a second set of identifiers associated with each glycomolecule of the second set of glycomolecules, and wherein the second search output comprises: a third set of identifiers associated with each glycomolecule of the third set of glycomolecules; a fourth set of identifiers associated with each glycomolecule of the fourth set of glycomolecules, and wherein identifying the conflicting subset of glycomolecules comprises: (a) determining an identifier in the first set of identifiers that is the same as an identifier in the fourth set of identifiers; or (b) determining an identifier in the second set of identifiers that is the same as an identifier in the third set of identifiers.

[0006] Also provided is a method of identifying glycomolecules in a biological sample, comprising: receiving a first search output for a biological sample, the first search output being associated with a first glycomolecule search engine; receiving a second search output for the biological sample, the second search output being associated with a second glycomolecule search engine; and determining that the first and second search outputs identify a conflicting subset of glycomolecules, wherein the conflicting subset of glycomolecules comprises one or more glycomolecules determined as being both of a first type in the first search output and of a second type in the second search output. Optionally, the one or more glycomolecules of the conflicting subset of glycomolecules are determined as not being of the second type in the first search output. In some embodiments, the one or more glycomolecules of the conflicting subset of glycomolecules are determined as not being of the first type in the second search output. In some embodiments, the first search output comprises a set of identifiers associated with each glycomolecule of a first set of glycomolecules, wherein the second search output comprises a second set of identifiers

associated with each glycomolecule of a second set of glycomolecules, and wherein determining that the first and second search outputs identify the conflicting subset of glycomolecules comprises: identifying a first identifier in the first set of identifiers associated with a glycomolecule of the first type in the first set of glycomolecules; and identifying a second identifier in the second set of identifiers associated with a glycomolecule of the second type in the second set of glycomolecules, wherein the first and second identifiers are the same.

[0007] In some embodiments, the method further includes generating a notification comprising a listing of the conflicting subset of glycomolecules. In some embodiments, the method further includes: selecting a subset of data corresponding to the conflicting subset of glycomolecules; and performing an additional analysis on the selected data to confirm an identity of the one or more glycomolecules of the conflicting subset of glycomolecules.

[0008] In some embodiments, the first search output comprises a first set of confidence scores associated with the first or the second set of glycomolecules, and the second search output comprises a second set of confidence scores associated with the third or the fourth set of glycomolecules. Optionally, each confidence score of the first set of confidence scores is equal to or greater than a first confidence score threshold, and each confidence scores of the second set of confidence scores is equal to or greater than a second confidence score threshold.

[0009] In some embodiments, the method further includes resolving the conflicting subset of glycomolecules based on a comparison of confidence scores of the first and second sets of confidence scores. Optionally, resolving the conflicting subset of glycomolecules comprises, for each glycomolecule identified as being present in both the first set and the fourth set, or in both the second set and the third set, determining that the glycomolecule is of the type identified by the search output comprising the higher confidence score associated with the glycomolecule. In some embodiments, the first and second sets of confidence scores are normalized.

[0010] In some embodiments, glycomolecules of the first type and glycomolecules of the second type have masses within about 0-8 daltons of each other. In some embodiments, the first type comprises a sialic acid containing glycomolecule and the second type comprises a fucosylated glycomolecule. In some embodiments, the first search output is in a first format, wherein the second search output is in a second format, and wherein the method further comprises, converting one or both of the first and second search

outputs to a common format, wherein the common format comprises the first format, the second format, or a universal glycomolecular data format.

[0011] Also provided is a method of determining a glycomolecule profile of a biological sample, comprising: receiving a first search output for a biological sample, the first search output being associated with a first glycomolecule search engine, wherein the first search output comprises a first set of glycomolecules associated with a first set of confidence scores; receiving a second search output for the biological sample, the second search output being associated with a second glycomolecule search engine, wherein the second search output comprises a second set of glycomolecules associated with a second set of confidence scores; and determining a glycomolecule profile of the biological sample based on (1) an overlap between the first and second search outputs and (2) the first and second set of confidence scores. Optionally, the first search output comprises a first set of identifiers, and the second search output comprises a second set of identifiers, and wherein determining the glycomolecule profile further comprises: determining the first set of identifiers associated with the first set of glycomolecules; and determining the second set of identifiers associated with the second set of glycomolecules. In some embodiments, the overlap comprises a set of glycomolecules in both the first and second search outputs, wherein each glycomolecule in the first set of glycomolecules is associated with a first confidence score of the first set of confidence scores that is equal to or greater than a first confidence score threshold, and each glycomolecule in the second set of glycomolecules is associated with a second confidence score of the second set of confidence scores that is equal to or greater than a second confidence score threshold. Optionally, the method further includes setting the value of the first confidence score threshold and the value of the second confidence score threshold to optimize a percentage overlap between the search outputs among all the first and second glycomolecules having, respectively, first confidence scores equal to or greater than the first confidence score threshold and second confidence scores equal to or greater than the second confidence score threshold. In some embodiments, determining the glycomolecule profile is based on an overlap between the first set of glycomolecules having first confidence scores within a first highest percentage and the second set of glycomolecules having second confidence scores within a second highest percentage.

[0012] In some embodiments, the method further includes generating a consensus list of glycomolecules based on the overlap. Optionally, generating the

consensus list comprises applying a weighting factor to either one or both of the first and second sets of confidence scores.

[0013] In some embodiments, the first search output is in a first format, and wherein the second search output is in a second format, wherein the method further comprises, before determining the glycomolecule profile, converting one or more of the first and second search outputs to a common format, wherein the common format comprises the first format, the second format, or a universal glycomolecular data format. Optionally, the common format is the universal glycomolecular data format. In some embodiments, the overlap comprises one or more of peptide sequences, protein IDs, retention times, m/z values, or charge states.

[0014] In some embodiments, the first search output and second search output are generated by running the respective glycomolecule search engines on the same glycomolecule mass spectrometry data of the biological sample.

[0015] In some embodiments, the glycomolecule is a glycopeptide, glycoprotein, glycolipid, glycoRNA, or glycoDNA. In some embodiments, the glycomolecule search engine is a glycopeptide search engine, glycoprotein search engine, glycolipid search engine, glycoRNA search engine, or glycoDNA search engine. In some embodiments, the biological sample is from human urine, human stool, human blood, or human serum.

[0016] Also provided is an electronic system comprising: a processor; and non-transitory memory comprising instructions, which when executed by the processor causes the processor to perform any of the method of the present disclosure.

[0005] Provided herein is a system for curating a consensus list of glycomolecules identified in a biological sample on a user interface, the system comprising: a processor; and a memory comprising instructions that, when executed by the processor, causes the processor to: retrieve a plurality of glycomolecule search results sets for a biological sample from a plurality of glycomolecule search engines, wherein each glycomolecule search results set comprises glycomolecule search results identifying one or more predicted glycomolecules as being present, wherein each predicted glycomolecule is associated with a unique identifier; determine a consensus set comprising, for each unique identifier, a consensus search result identifying one of the predicted glycomolecules associated with the unique identifier in the glycomolecule search results sets; provide a user interface configured to display: one or more search results elements configured to display one or more representations of the one or more predicted glycomolecules from each of the

plurality of glycomolecule search results sets; and a consensus list element configured to display the consensus set. Optionally, representations corresponding to the glycomolecule search results for predicted glycomolecules associated with the same unique identifiers are displayed in each of the search result elements. Optionally, the user interface is further configured to display the unique identifier in association with the one or more representations. Optionally, the consensus set is determined at least in part by: identifying conflicting search results and non-conflicting search results among the glycomolecule search results; and populating the consensus list element with the non-conflicting search results.

[0006] In some embodiments, the glycomolecule search results are further evaluated by: resolving the conflicting search results by applying a conflict resolution rule set to generate resolved conflicting search results; and populating the consensus list element with the resolved conflicting search results. Optionally, the conflict resolution rule set comprises identifying, for a particular conflicting search result associated with a particular unique identifier, one or more plurality glycomolecules from the predicted glycomolecules of the conflicting search results associated with the particular unique identifier.

[0007] In some embodiments, the search results comprise a confidence score for each of the glycomolecules in the conflict search results, and wherein the conflict resolution rule set comprises identifying a consensus search result based on a comparison of the confidence scores of each search result among the conflict search results. In some embodiments, determining the consensus set comprises determining an initial consensus set, wherein the initial consensus set comprises a single predicted glycomolecule for each unique identifier.

[0008] In some embodiments, the instructions, when executed by the processor, further causes the processor to provide the plurality of search result elements, each search result element configured to receive recognize a user input to select one predicted glycomolecule among the predicted glycomolecules associated with the same unique identifier as the consensus search result for the unique identifier. Optionally, the user input comprises a touch and a scrolling activity, and wherein the user interface is further configured to scroll through the representations of the predicted glycomolecules displayed in the search results element in response to the scrolling activity.

[0009] In some embodiments, the user interface is configured to arrange the representations of the predicted glycomolecules of each search results set with a respective search results element into an ordered list for each of the search results set. Optionally, the

user interface is further configured to scroll through the ordered lists in a coordinated manner in response to a scrolling activity on one of the ordered lists by the user. Optionally, each of the ordered lists is identically ordered.

[0010] In some embodiments, the unique identifier is a mass spectrometry scan number, and wherein the mass spectrometry scan number identifies a glycomolecule query in each of the glycomolecule search engines.

[0011] In some embodiments, the system further comprises a human interface device. In some embodiments, the system comprises a display. Optionally, the display further comprises a sensor. Optionally, the display further comprises an actuator. In some embodiments, the user interface is configured to: recognize an input of a user; recognize a selected glycomolecule when the input of the user hovers or interacts with the glycomolecule shown in the glycomolecule search results sets or the consensus set; and present a visualized glycomolecule for the selected glycomolecule. Optionally, the visualized glycomolecule comprises one or more potential structures of the selected glycomolecule. In some embodiments, the visualized glycomolecule is shown in a modal window in the user interface. In some embodiments, the visualized glycomolecule is shown in a popup window in the user interface. In some embodiments, the visualized glycomolecule is shown in a dedicated region of the user interface. In some embodiments, the user interface is configured to: recognize an input of a user and permit the user to manually edit one or more of the consensus search result; and optionally permit the user to drag and drop one or more of the predicted glycomolecules from one or more of the glycomolecule search results sets into the consensus set. In some embodiments, the user interface is configured to: recognize an input of a user and permit the user to drag and drop one or more of the predicted glycomolecules from one or more of the glycomolecule search results sets into the consensus set.

[0012] In some embodiments, the glycomolecules comprise glycans, glycoproteins, glycolipids, glycoRNAs, glycoDNAs, or any combination thereof. In some embodiments, the glycomolecules comprise glycans.

[0013] In some embodiments, the instructions, when executed by the processor, further causes the processor to convert one or more of the glycomolecule search results to a common format. Optionally, the common format is a universal glycan format. Optionally, the universal glycan format comprises: a first number *l* corresponding to an amount of hexose; a second number *m* corresponding to an amount of N-acetylhexosamine; a third number *n* corresponding to an amount of fucose; and a fourth number *o* corresponding to

an amount of N-acetylneuraminic acid, and wherein *l*, *m*, *n*, *o* are each an integer ≥ 0. In some embodiments, the glycomolecule search results from each of the glycomolecule search engines is in a platform-specific format.

[0014] Also provided herein is a method for curating a consensus list of glycomolecules identified in a biological sample on a user interface, comprising: at a computer system including one or more processors, and memory storing one or more programs configured to be executed by the one or more processors, the one or more programs including instructions for: retrieving a plurality of glycomolecule search results sets for a biological sample from a plurality of glycomolecule search engines wherein each glycomolecule search results set comprises glycomolecule search results identifying one or more predicted glycomolecules as being present, wherein each predicted glycomolecule is associated with a unique identifier; determining a consensus set, wherein the consensus set comprises, for each unique identifier, a consensus search result identifying one of the predicted glycomolecules associated with the unique identifier in the glycomolecule search results sets; providing a user interface configured to display: one or more search results elements configured to display one or more representations of the one or more predicted glycomolecules from each of the plurality of glycomolecule search results sets; and a consensus list element configured to display the consensus set. Optionally, representations corresponding to the glycomolecule search results for predicted glycomolecules associated with the same unique identifiers are displayed in each of the search result elements. Optionally, the user interface is further configured to display the unique identifier associated with the one or more representations.

[0015] In some embodiments of the method, the determining the consensus list step comprises: identifying conflicting search results and non-conflicting search results among the glycomolecule search results; and populating the consensus list element with the non-conflicting search results. Optionally, the one or more programs further including instructions for: resolving the conflicting search results to generate resolved conflicting search results; and populating the consensus list element with the resolved conflicting search results. Optionally, the conflict resolution rule set comprises identifying, for a particular conflicting search result associated with a particular unique identifier, one or more plurality glycomolecules from the predicted glycomolecules of the conflicting search results associated with the particular unique identifier. Optionally, the conflict resolution rule set further comprises identifying a consensus search result based on relative confidence scores of each search result among the conflict search results.

[0016] In some embodiments of the method, the consensus set comprises an initial consensus set, and wherein the initial consensus set comprises a single search predicted glycomolecule for each unique identifier. In some embodiments of the method, the one or more programs further include instructions for providing the plurality of search result elements, each of the search result elements configured to recognize a user input to select one predicted glycomolecule among the predicted glycomolecules associated with the glycomolecule search engine that is associated with the same unique identifier. In some embodiments of the method, the one or more programs further include instructions for arranging, in the user interface, the representations of the predicted glycomolecules of each search results set with a respective search results element into an ordered list for each of the search results set. Optionally, the one or more programs further include instructions for ordering each of the ordered lists identically.

[0017] In some embodiments of the method, the input comprises a touch and a scrolling activity, and wherein the user interface is further configured to scroll through the ordered lists at about a same scrolling speed and about a same scrolling direction in response to the scrolling activity on one of the ordered lists. In some embodiments of the method, the unique identifier is a mass spectrometry scan number, and wherein the mass spectrometry scan number identifies a glycomolecule query in each of the glycomolecule search engines.

[0018] In some embodiments of the method, the user interface is configured to: recognize an input of a user and permit the user to manually edit one or more of the consensus search result; and optionally permit the user to drag and drop one or more of the predicted glycomolecules from one or more of the glycomolecule search results sets into the consensus set. In some embodiments of the method, wherein the user interface is configured to: recognize an input of a user and permit the user to drag and drop one or more of the predicted glycomolecules from one or more of the glycomolecule search results sets into the consensus set. In some embodiments of the method, the search results further comprise a confidence score for each of the glycomolecules in the conflict search results, wherein the user interface is configured to: recognize an input of a user and permit the user to apply a custom confidence score to any of the glycomolecules in the conflict search results. Optionally, the one or more programs further include instructions to update a machine learning process with the custom confidence scores, wherein the machine learning process is configured to output recommendations and predictions.

[0019] In some embodiments of the method, the user interface is configured to: recognize an input of a user; recognize a selected glycomolecule when the input of the user hovers or interacts with the glycomolecule shown in the glycomolecule search results sets or the consensus set; and present a visualized glycomolecule for the selected glycomolecule. Optionally, the visualized glycomolecule comprises one or more potential structures of the selected glycomolecule. In some embodiments of the method, the visualized glycomolecule is shown in a modal window in the user interface. In some embodiments of the method, the visualized glycomolecule is shown in a popup window in the user interface. In some embodiments of the method, the visualized glycomolecule is shown in a dedicated region of the user interface.

[0020] In some embodiments of the method, the glycomolecules comprise glycans, glycolipids, glycoRNAs, glycoDNAs, or any combination thereof. Optionally, the representation of the at least one glycomolecule comprises a glycan representation in a universal glycan format. In some embodiments of the method, the glycomolecules comprise glycans. Optionally, the representation of the at least one glycomolecule comprises a glycan representation in a universal glycan format.

[0021] In some embodiments, the universal glycan format comprises: a first number *l* corresponding to an amount of hexose; a second number *m* corresponding to an amount of N-acetylhexosamine; a third number *n* corresponding to an amount of fucose; and a fourth number *o* corresponding to an amount of N-acetylneuraminic acid, and wherein *l*, *m*, *n*, *o* are each an integer ≥ 0.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] Fig. 1 is a block diagram which illustrates a glycan database input converter connected to a series of glycan databases connected to various search engines along with a glycopeptide data converter that converts the output of each search engine to a universal glycan format, according to some non-limiting embodiments of the present disclosure.

[0020] Fig. 2 is a schematic diagram which illustrates the conversion of a particular universal glycan format into the different input formats found within search engine-specific glycan databases, according to some non-limiting embodiments of the present disclosure.

[0021] Fig. 3 is a schematic diagram showing direct conversions of glycan representations between different glycan formats that are each compatible with a glycomolecule search engine.

[0022] Fig. 4 is a schematic diagram showing conversion of glycan representations between different glycan formats via a universal glycan format, according to some non-limiting embodiments of the present disclosure.

[0023] Fig. 5 is a block diagram showing a method of converting a glycan data format, according to some non-limiting embodiments of the present disclosure.

[0024] Fig. 6 is a block diagram showing a method of converting a glycan data format, according to some non-limiting embodiments of the present disclosure.

[0025] Fig. 7 is a block diagram showing a computer system for implementing a glycan data management system and a method of converting a glycan data format, according to some non-limiting embodiments of the present disclosure.

[0026] Fig. 8 is a block diagram showing a network system for implementing a glycan data management system and a method of converting a glycan data format, according to some non-limiting embodiments of the present disclosure.

[0027] Fig. 9 is a schematic diagram showing conversion of glycan representations between different glycan formats via a universal glycan format, according to some non-limiting embodiments of the present disclosure.

[0028] Fig. 10 is a table showing glycan representations in different glycan formats, including a universal glycan format, according to some non-limiting embodiments of the present disclosure.

[0022] Fig. 11 is a schematic diagram showing comparison of search outputs from different glycomolecule search engines using a common format, according to some non-limiting embodiments of the present disclosure.

[0023] Fig. 12 is a schematic diagram showing a method of comparing glycomolecule search outputs from two or more glycomolecule search engines, according to some non-limiting embodiments of the present disclosure.

[0024] Fig. 13 is a schematic diagram showing a method for analyzing a biological sample to determine a glycomolecule profile, according to some non-limiting embodiments of the present disclosure.

[0025] Fig. 14 is a block diagram showing a computer system for implementing a glycomolecule data analysis system and a method of comparing glycomolecule search outputs, according to some non-limiting embodiments of the present disclosure.

[0026] Fig. 15 is a block diagram showing a network system for implementing a glycomolecule data analysis system and a method of comparing glycomolecule search outputs, according to some non-limiting embodiments of the present disclosure.

[0027] Fig. 16 is a collection of tables showing portions of search outputs from a PGLYCO database search and a BYONIC database search run on the same glycopeptide mass spectrometry data, each output having a search engine-specific format.

[0028] Fig. 17 is a collection of tables showing a search output from a glycopeptide search engine in a search engine-specific format and a universal glycomolecular data format, where the Scan number of the PGLYCO search result is matched to the MS2 protein number of the universal glycopeptide format, according to some non-limiting embodiments of the present disclosure.

[0029] Fig. 18 is a collection of tables showing a search output from a glycopeptide search engine in a search engine-specific format and a universal glycomolecular data format, according to some non-limiting embodiments of the present disclosure.

[0030] Fig. 19 is a collection of tables showing a search output from a glycopeptide search engine in a search engine-specific format and a universal glycomolecular data format, where the Scan number of the METAMORPHEUS search result is matched to the MS2 protein number of the universal glycopeptide format, according to some non-limiting embodiments of the present disclosure.

[0031] Fig. 20 is a schematic diagram showing extraction of glycopeptide data parameters for an identified glycopeptide in a search output of a glycopeptide search engine run on glycopeptide mass spectrometry data, according to some non-limiting embodiments of the present disclosure.

[0032] Fig. 21 is a schematic diagram showing extraction of glycopeptide data parameters for an identified glycopeptide in a search output of a glycopeptide search engine run on glycopeptide mass spectrometry data, according to some non-limiting embodiments of the present disclosure.

[0033] Fig. 22 is a schematic diagram showing extraction of glycopeptide data parameters for an identified glycopeptide in a search output of a glycopeptide search engine run on glycopeptide mass spectrometry data, according to some non-limiting embodiments of the present disclosure.

[0017] FIG. 23 is a schematic diagram showing a method of identifying glycomolecules in a biological sample, according to some non-limiting embodiments of the present disclosure.

[0018] FIG. 24A is a schematic diagram showing a method of identifying glycomolecules in a biological sample, according to some non-limiting embodiments of the present disclosure.

[0019] FIGs. 24B-24D illustrate examples of different pairs of glycans with similar masses. Legend: Gal: D-Galactose; Man: D-Mannose; Fuc: L-Fucose; GlcNAc: N-Acetyl-D-Glucosamine; Neu5Ac: N-Acetylneuraminic Acid; Neu5Gc: N-Glycolylneuraminic acid.

[0020] FIG. 25 is a schematic diagram showing a method of determining a glycomolecule profile of a biological sample, according to some non-limiting embodiments of the present disclosure.

[0021] FIG. 26 is a block diagram showing a computer system for implementing a method of the present disclosure, according to some non-limiting embodiments of the present disclosure.

[0022] FIG. 27 is a block diagram showing a network system for implementing a method of the present disclosure, according to some non-limiting embodiments of the present disclosure.

[0023] FIGs. 28A and 28B are a collection of diagrams showing overlapping and no-overlapping sets of glycomolecules identified as containing sialic acid or as being fucosylated by two different glycopeptide search engines, according to some non-limiting embodiments of the present disclosure.

[0024] FIGs. 29A and 29B are a collection of diagrams showing the change in the overlapping set of glycomolecules identified by two different glycopeptide search engines based on search engine-specific confidence scores, when the confidence score thresholds ("Cut-off score") are altered in a search engine-specific manner, according to some non-limiting embodiments of the present disclosure.

[0022] FIG. 30 is a schematic diagram showing a non-limiting example, according to some embodiments of the disclosure, of a user interface presenting glycomolecule search results displayed in search result elements and a consensus set displayed in a consensus list.

[0023] FIG. 31A is a schematic diagram showing a non-limiting example, according to some embodiments of the disclosure, of dynamic interaction with a user interface presenting glycomolecule search results displayed in search result elements and a consensus set displayed in a consensus list.

[0024] FIG. 31B is a schematic diagram showing a non-limiting example, according to some embodiments of the disclosure, of a user interface that presents a glycomolecule visualization.

[0025] FIG. 31C is a schematic diagram showing a non-limiting example, according to some embodiments of the disclosure, of a user interface presenting glycomolecule search results displayed in search result elements and a consensus set displayed in a consensus list.

[0026] FIG. 32 is a schematic diagram showing a non-limiting example, according to some embodiments of the disclosure, of a user interface presenting glycomolecule search results displayed in search result elements and a consensus set displayed in a consensus list.

[0027] FIG. 33 is a schematic diagram showing a non-limiting example, according to some embodiments of the disclosure, of a user interface presenting glycomolecule search results displayed in search result elements and a consensus set displayed in a consensus list.

[0028] FIG. 34 is a block diagram showing a non-limiting example of a computer system implementing systems and methods of the disclosure.

[0029] FIG. 35 is a block diagram showing a non-limiting example of a method for populating a consensus list element according to some non-limiting embodiments of the disclosure.

[0030] FIG. 36 is a schematic diagram showing conversion of glycan representations between different glycan formats via a universal glycan format, according to some non-limiting embodiments of the disclosure.

[0031] FIG. 37 is a schematic diagram showing conversion of glycan representations between different glycan formats via a universal glycan format, according to some non-limiting embodiments of the disclosure.

[0029] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

DETAILED DESCRIPTION

[0030] The following description is presented to enable one of ordinary skill in the art to make and use the invention and to incorporate it in the context of particular applications. Various modifications, as well as a variety of uses in different applications

will be readily apparent to those skilled in the art, and the general principles defined herein may be applied to a wide range of embodiments. Thus, the inventions herein are not intended to be limited to the embodiments presented, but are to be accorded their widest scope consistent with the principles and novel features disclosed herein.

[0031] All the features disclosed in this specification, (including any accompanying claims, abstract, and drawings) may be replaced by alternative features serving the same, equivalent or similar purpose, unless expressly stated otherwise. Thus, unless expressly stated otherwise, each feature disclosed is one example only of a generic series of equivalent or similar features.

[0032] Please note, if used, the labels left, right, front, back, top, bottom, forward, reverse, clockwise and counterclockwise have been used for convenience purposes only and are not intended to imply any particular fixed direction. Instead, they are used to reflect relative locations and/or directions between various portions of an object.

I. OVERVIEW

[0033] In some embodiments, a user may introduce a sample into an analyzer device to determine identities or properties of analytes within the sample. For example, the user may introduce a biological sample including glycomolecules into a mass spectrometer. The user may then initiate a multiple reaction monitoring mass spectrometry (MRM-MS) process. The MRM-MS process may include ionizing glycomolecule analytes (precursor ions) within the sample, fragmenting the ionized analytes (product ions), and measuring charge signals from the precursor and/or product ions. These measured signals may be used to identify the analytes and determine relative abundances of the analytes in the sample. This identification and determination may include querying one or more glycan databases (along with other databases, e.g., protein databases in the case of glycopeptides) by inputting the measured signals into an appropriate search engine.

[0034] In some embodiments, as shown in Fig. 1, the system includes a first Glycan Database Converter, which may be a glycan format converter, which is configured to convert naming formats or representations used in a particular glycan database to a different naming format or representation. Currently, there exist a number of different glycan databases that identify or represent glycans by their composition and/or structure, but many of these glycan databases use different formats to identify or represent these glycans. Although these various formats describe glycan composition and/or structure, their formats are uniquely tailored to fit the conventions of associated glycan/glycopeptide

search engines, which each operate based on their own respective glycan input formats and associated glycan databases. As shown in Fig. 1, the Glycan Database Converter communicates with various search engine databases. In addition, other public glycan databases may also communicate with the Glycan Database Converter. Each of the glycan databases may communicate and provide input to the various search engines as shown in Fig. 1. In some embodiments, search results from each of the search engines may then be outputted and directed into a Glycomolecule Search Converter (e.g., a Glycomolecule Search Converter), which is configured to obtain the specific output formats from each search engine and convert it into a universal glycomolecule format so that the results from each search can be compared to one another.

[0035] Some examples of a conversion are illustrated in Fig. 2, which illustrates how a glycan representation can be converted between a universal glycan format (e.g., N(3)H(3)F(1)A(0)) and various platform-specific formats compatible with particular search engines (e.g., PGLYCO3, BYONIC, and METAMORPHEUS search engines).

[0036] Fig. 3 illustrates an example method of converting directly between platform-specific formats. As illustrated, direct conversion of the first format **310** of the glycan data in a database of interest (e.g., that is compatible with a first search engine) to second and third formats **320**, **330** compatible with one or more other search engines requires conversion from the first format to each of the other formats **315**, **325**. If conversion of glycan data formats from the second and third formats is desired (e.g., to use a glycan database for a second search engine that takes in glycan data in the second format on a third search engine that takes in glycan data in the third format, but not the second format), another conversion **345** is required. With the addition of each N^{th} new glycan data format **340**, where N is an integer greater than 1, the number of conversion rules (e.g., for converting each of the existing glycan formats to the new format (and vice versa) **335**, **355**, **365** increases by N–1. In general, the total number of conversion rule sets for converting glycan data formats among all of the N glycan data formats (counting each pair-wise conversion in both directions as a single conversion rule set) will be N*(N–1)/2. As discussed herein, embodiments of the present disclosure can reduce the number of conversion rule sets required to convert glycan data formats across different glycan databases.

[0037] In some embodiments, the disclosed glycan format converter may be configured to convert among the different formats so that the user can search each of the different glycan databases using any desired search engine. Additionally or alternatively,

the glycan format converter may be configured to convert any of the different formats into a newly developed "universal" format as described herein, so that users can rely on this single universal glycan format rather than having to convert and otherwise work with multiple different formats. Similarly, the glycan format converter may also be configured to convert from this universal glycan format to any other suitable or compatible format (e.g., the format associated with BYONIC, PGLYCO3, or any other suitable search engine). Without being limited by theory, when a common, universal glycan format is used for conversion between different formats, for each new format, only one new conversion rule is needed to convert between the new format and the universal glycan format, and the new format can then be converted to any of the other pre-existing formats. In such cases, the total number of conversion rules for converting glycan data formats among all of the N glycan data formats will be reduced as compared to a case where a universal glycan format is not used in the conversion process. For example, the total number of conversion rules in such cases would be N (rather than N*(N–1)/2), where N is an integer greater than 1.

[0038] In some embodiments, the glycan format converter may be capable of converting between structural formats (such as the PGLYCO3 format), which indicate the structure of a glycan, and purely compositional formats (such as the Metamorpheus format), which may indicate only the chemical composition of a glycan.

[0040] In some embodiments, the system may include a second converter, which may be a glycomolecule search result format converter (e.g., a glycopeptide search result format converter). Different search engines output the results of each search in a different glycopeptide format (see, for example, Fig. 16). For example, the BYONIC search engine reports the amino acid sequence both before and after the peptide structure (see, e.g., Fig. 16, bottom panel). In comparison, the PGLYCO3 search engine reports the actual glycosylation site as amino acid "J" instead of the real amino acid "N", which is asparagine (see, e.g., Fig. 16, top panel). As another example, a BYONIC search output may represent a glycan of a glycomolecule as "HexNAc(3)Hex(3)Fuc(1)," while a PGLYCO 3 search output may represent the same glycan as "(N(F) (N(H(H) (H(N))))) where N corresponds to HexNAc and H corresponds to Hex."

[0041] The methods and systems disclosed herein enable the running of searches on sample data (e.g., mass spectrometry data) across multiple glycomolecule search engines and comparison of the search outputs. Each glycomolecule search engine may have its own unique algorithms for predicting or determining properties or identities of analytes in a sample, such that different search engines may arrive at different

conclusions of varying confidence levels about the properties or identities of the same set of analytes. In some cases, identification of glycomolecules by one search engine may be in conflict with identification by another search engine. By comparing the search results, a user can find potentially mis-identified glycomolecules in the search results. Closer analysis of the underlying data may allow resolution of the conflicting identification.

[0041] Relatedly, some search engines may be better for certain types of searches than other search engines. Comparing search outputs for the same sample data across multiple search engines can provide a more complete picture, can help validate or corroborate the identity and properties of analytes, and can generally increase the confidence level a researcher, clinician, or other user has in making conclusions based on the search outputs. Comparing search outputs for the same sample data across multiple search engines may provide guidance as to the pros and cons of different search engines. For example, systematic mis-identification of glycomolecules by one search engine may reveal biases in the search algorithm used by the search engine. Additional routines may then be developed to account for such biases for analyzing future searches.

[0042] Relying on multiple search engines may also provide a more comprehensive annotation of the sample data. However, if a user searches the same sample data across multiple search engines and attempts to compare the search outputs, the varying formats in the search outputs of the different search engines make this comparison difficult. The methods and systems disclosed herein can, at least in part, enable conversion of search outputs to a common format (e.g., to a format specific to one of the search engines, a universal glycomolecular data format) in order to facilitate the comparison process. Although this disclosure focuses on comparing searches based on mass spectrometry data to predict/determine properties or identities of glycomolecules, the disclosure contemplates comparing searches based on any sample data suitable for predicting/determining properties or identities of glycomolecules.

[0042] Using the glycopeptide search result format converter, researchers can directly compare their search results despite using multiple search engines because each search result will be converted from the specific format output by each search engine into a common format, e.g., a universal format. Moreover, the system can correlate and manage the different confidence scores that are reported for each search results from the individual search engines, to provide a system for comparing the confidence of each result across the databases. These correlations can be determined by, in part, comparing the search results of multiple engines to determine the overlap, and then deriving equivalent confidence

values. Thus, in some cases, the use of a common format, such as a universal format, to which output from different glycomolecule search engines are converted (e.g., as provided in the present disclosure), allows a user to more easily compare the search results for the same experiment or mass spectrometry run.

[0032] Embodiments provide options to correlate and manage the different confidence scores that are reported for each search result from the individual search engines, to compare the confidence that each respective search engine has in the identity of the search result. These correlations can be determined by, in part, comparing the search results of multiple search engines to determine the overlap, and then deriving equivalent confidence values.

[0039] Certain techniques for analyzing biological samples using mass spectroscopy are known. See, for example, International PCT Patent Application Publication No. WO2019079639A1, filed October 18, 2018 as International Patent Application No. PCT/US2018/56574, and titled IDENTIFICATION AND USE OF BIOLOGICAL PARAMETERS FOR DIAGNOSIS AND TREATMENT MONITORING, the entire contents of which are herein incorporated by reference in its entirety for all purposes. See also US Patent Application Publication No. US20190101544A1, filed August 31, 2018 as US Patent Application No. 16/120,016, and titled Identification and Use of Glycopeptides as Biomarkers for Diagnosis and Treatment Monitoring, the entire contents of which are herein incorporated by reference in its entirety for all purposes.

[0035] Disclosed herein are systems and methods that facilitate curation of a consensus list of glycomolecules identified in a biological sample by providing a user interface for displaying one or more glycomolecule search results from one or more glycomolecule search engines and the consensus list of the predicted glycomolecules. In one embodiment, the consensus list determined by one or more conflict resolution rules may be automatically populated by the system. Additionally, the systems and methods further facilitate curation of the consensus list by permitting manual curation thereof via user input, thereby allowing further determination and selection of the predicted glycomolecules in the consensus list. The presentation of search results and a consensus list on a user interface, according to embodiments of the present systems and methods, can allow a user to compare search results across different search engines and have increased confidence in the quality of the identity of glycomolecules present in a sample. In some embodiments, comparison of search results across different search engines may allow

identification of strengths and/or systematic biases of each search engine compared to others. Advantageously, in some embodiments, manual curation of the consensus list not only allows for independent assessment of the glycomolecules shown therein but also improves accuracy and streamlines the process of identifying glycomolecules in a biological sample, e.g., by using the manual curation to inform conflict resolution rules applied to future analyses.

[0036] While aspects of this disclosure relate to curation of a consensus list from glycomolecules identified in biological samples used in medical diagnostics and life sciences research, the disclosure is not limited to mass spectrometry data analysis, as the disclosure contemplates any other suitable systems or methods that utilize a comparative analysis of results from multiple sources as described in more detail below.

II. DEFINITIONS

[0040] As used herein, the singular forms "a," "an" and "the" include plural referents unless the context clearly dictates otherwise.

[0041] As used herein, the phrase "biological sample," refers to a sample derived from, obtained by, generated from, provided from, taken from, or removed from an organism, or from fluid or tissue from the organism. Biological samples include, but are not limited to synovial fluid, whole blood, blood serum, blood plasma, urine, sputum, tissue, saliva, tears, spinal fluid, tissue section(s) obtained by biopsy; cell(s) that are placed in or adapted to tissue culture; sweat, mucous, fecal material, gastric fluid, abdominal fluid, amniotic fluid, cyst fluid, peritoneal fluid, pancreatic juice, breast milk, lung lavage, marrow, gastric acid, bile, semen, pus, aqueous humor, transudate, and the like including derivatives, portions and combinations of the foregoing. In some examples, biological samples include, but are not limited, to blood and/or plasma. In some examples, biological samples include, but are not limited, to urine or stool. In some examples, biological samples include, but are not limited, to saliva. In some examples, biological samples include, but are not limited to, tissue dissections and tissue biopsies. In some examples, biological samples include, but are not limited to, any derivative or fraction of the aforementioned biological samples.

[0042] As used herein, the term "glycan" refers to the carbohydrate residue of a glycoconjugate, such as the carbohydrate portion of a glycopeptide, glycoprotein, glycolipid, or proteoglycan. Glycans can be monomers or polymers of sugar residues, but typically contain at least three sugars, and can be linear or branched. A glycan may include

natural sugar residues (e.g., glucose, N-acetylglucosamine, N-Acetylhexosamine (HexNAc), N-acetylneuraminic acid (NeuAc), galactose, mannose, fucose, hexose (Hex) , arabinose, ribose, xylose, etc.) and/or modified sugars (e.g., 2′-fluororibose, 2′-deoxyribose, phosphomannose, 6′-sulfo N-acetylglucosamine, etc). The term "glycan" includes homo and heteropolymers of sugar residues. The term encompasses free glycans, including glycans that have been cleaved or otherwise released from a glycoconjugate. Glycan structures (as compared to glycan data formats or representations) are described by a glycan reference code number, and also illustrated in International PCT Patent Application No. PCT/US2020/016286, filed January 31, 2020, which is herein incorporated by reference in its entirety for all purposes.

[0043] "Glycomolecule" as used herein includes glycans and glycoconjugates (such as, but not limited to, glycopeptides, glycoproteins, glycolipids, glycoRNA, glycoDNA, etc.). Glycomolecule includes fragments of glycoconjugates. As used herein, the term "glycopeptide," refers to a peptide having at least one glycan residue bonded thereto. A glycopeptide can be an intact protein (e.g., a glycoprotein) or any fragment thereof that has at least one glycan residue covalently bonded thereto.

[0044] As used herein, the term "glycoform" refers to a unique primary, secondary, tertiary, and quaternary structure of a protein with an attached glycan of a specific structure.

[0045] As used herein, the phrase "glycosylated peptides," refers to a peptide bonded to a glycan. Glycosylated peptides include peptides that have been covalently modified by glycosylation to become bonded to a glycan.

[0046] As used herein, the phrase "glycopeptide fragment" or "glycosylated peptide fragment" or "glycopeptide" refers to a glycosylated peptide (or glycopeptide) having an amino acid sequence that is the same as part (but not all) of the amino acid sequence of the glycosylated protein (or glycoprotein) from which the glycosylated peptide (or glycopeptide) is obtained, e.g., ion fragmentation within a MRM-MS instrument. MRM refers to multiple-reaction-monitoring. Unless specified otherwise, within the specification, "glycopeptide fragments" or "fragments of a glycopeptide" refer to the fragments produced directly by using a mass spectrometer optionally after the glycoprotein has been digested enzymatically to produce the glycopeptides.

[0047] As used herein, the phrase "glycoprotein" refers to the glycosylated protein from which the glycosylated peptide is obtained. "Glycoprotein" refers to a protein that contains a peptide backbone covalently linked to one or more sugar moieties (i.e.,

glycans). As is understood by those skilled in the art, the peptide backbone typically comprises a linear chain of amino acid residues. The sugar moiety(ies) may be in the form of monosaccharides, disaccharides, oligosaccharides, and/or polysaccharides. The sugar moiety(ies) may comprise a single unbranched chain of sugar residues or may comprise one or more branched chains. In certain embodiments, sugar moieties may include sulfate and/or phosphate groups. Alternatively or additionally, sugar moieties may include acetyl, glycolyl, propyl or other alkyl modifications. In certain embodiments, glycoproteins contain O-linked sugar moieties; in certain embodiments, glycoproteins contain N-linked sugar moieties.

[0048] As used herein, the phrase "peptide," is meant to include glycopeptides unless stated otherwise.

[0049] As used herein, the phrase "multiple reaction monitoring mass spectrometry (MRM-MS)," refers to a highly sensitive and selective method for the targeted quantification of glycans and/or glycomolecules in biological samples. Unlike traditional mass spectrometry, MRM-MS is highly selective (targeted), allowing researchers to fine tune an instrument to specifically look for certain peptides' fragments of interest. MRM allows for greater sensitivity, specificity, speed, and quantitation of peptide fragments of interest, such as a potential biomarker. MRM-MS involves using one or more of a triple quadrupole (QQQ) mass spectrometer and a quadrupole time-of-flight (qTOF) mass spectrometer.

[0050] As used herein, the phrase "multiple-reaction-monitoring (MRM) transition," refers to the mass to charge (m/z) peaks or signals observed when a glycomolecule, or a fragment thereof, is detected by MRM-MS. The MRM transition is detected as the transition of the precursor and product ions.

[0051] "Representation" or "format" as used herein with reference to a glycan refers to any linear string of characters intended to convey compositional and/or structural features of a glycan. A glycan representation can be a string that includes symbols and/or alphanumerical characters. In some cases, a glycan representation or format can be constructed using pre-defined rules for representing the compositional and/or structural features of a glycan. For example, a glycan representation in an example format is N(3)H(3)F(1)A(0)G(0). This glycan representation indicates a glycan composed of 3 N-acetyl Acetylhexosamine molecules, 3 hexose molecules, 1 fucose molecule, and 0 N-acetylneuraminic acid molecules.

[0052] As used herein, "platform-specific glycan format" refers to any glycan format that is associated with one or more specific glycomolecule search engines, e.g., one or more specific glycomolecule search engines. A platform-specific glycan format can be used by, or be compatible with, the glycomolecule search engine. In some cases, the platform-specific glycan format is a conventional glycan format used by, or compatible with, conventional glycan databases and/or conventional glycan or glycopeptide search engines. Non-limiting examples of conventional glycan formats include formats used by PGLYCO3, BYONIC and METAMORPHEUS.

[0053] "Conversion rule" as used herein with reference to a glycan format refers to any step-wise operation for changing a glycan representation in one format to a different format, e.g., when executed by a processor on an input glycan data. A "conversion rule set" as used herein with reference to a glycan format refers to a collection of conversion rules that provide the step-wise operations for changing a glycan representation in one format to and from another format.

[0054] "Search engine" refers to any computer-implemented program configured to receive a query (e.g., an input string) and implement algorithms to identify entries within one or more databases that provide a match to the query that meets certain predefined criteria and/or user-specified criteria. A search engine is typically associated with its own proprietary glycan database and can rely on one or more statistical tests to determine the quality of any given match, and may provide a confidence score that reflects the quality of a match. A "glycomolecule search engine" refers to any search engine for identifying one or more of glycans, glycopeptides, glycolipids, glycoRNA, glycoDNA, etc., in sample data (e.g., mass spectrometry sample data). Non-limiting examples of glycopeptide search engines include PGLYCO3, BYONIC and METAMORPHEUS. A platform-specific glycan format can include any glycan format that can be converted to and/or from a universal glycan format, as disclosed herein.

[0055] As used herein, the term "about" indicates and encompasses an indicated value and a range above and below that value. In certain embodiments, the term "about" indicates the designated value ± 10%, ± 5%, or ± 1%. In certain embodiments, the term "about" indicates the designated value ± one standard deviation of that value.

III. DESCRIPTION OF SYSTEMS AND METHODS

[0056] As mentioned above, there are a number of glycomolecule search engines available, including, for example, BYONIC, PGLYCO3, and METAMORPHEUS.

Comparison of search results from different glycomolecule search engines can inform the user about the reliability of the search results and about characteristics of a glycomolecule search engine. In some embodiments, the search results from different search engines can be compared between each other to identify conflicting subsets of glycomolecules identified in the search results. For example, one search engine may identify an analyte as being associated with one type of glycan (e.g., fucose), while another search engine may identify the same analyte as being associated with another type of glycan (e.g., sialic acid). In some embodiments, the search results from different search engines can be compared between each other to create a consensus list of identified glycomolecules.

[0057] As shown in Fig. 1, a universal glycan format can be used as an input to a Glycan Database Converter. The Glycan Database Converter can convert the format of a glycan database between a glycomolecule search engine-specific format and a universal glycan format, and/or between glycomolecule search engine-specific formats via the universal glycan format. Fig 2. shows a non-limiting example of taking a universal format for a molecule designated as "N(3)H(3)F(1)A(0)" and converting that universal format into PGLYCO 3, BYONIC and METAMORPHEUS formats. Thus, one search in the universal format can be entered into the system which will then be able to be searched across the various glycan databases. Although this disclosure gives example of the universal format designating the different sugars in a particular order, the disclosure contemplates any suitable order. For example, the order may be "H(3)N(3) F(1)A(0)." Similarly, the disclosure contemplates any number of sugars in a glycan representation in the universal format. For example, the glycan representation may include types of sugars (e.g., H(3)N(3)F(1)A(0)G(1)).

[0057] With reference to Fig. 4, non-limiting embodiments of glycan data management systems and methods for converting a glycan data format are provided. A glycan data management system can include a non-transitory memory that contains instructions for converting among a plurality of glycan formats (such as a plurality of platform-specific glycan formats, e.g., **410**, **420**, **430**, **440**) and a universal glycan format **400** by applying at least one of a plurality of format conversion rule sets (e.g., **415**, **425**, **435**, **445**). The plurality of format conversion rule sets can include at least a first rule set **415** configured to convert glycan representations of a first platform-specific glycan format **410** to and from the universal glycan format **400**; and a second rule set **425** configured to convert glycan representations of a second platform-specific glycan format **420** to and from

the universal glycan format **400**. The glycan data management system can further include a processor, e.g., a central processing unit (CPU), configured to: access an input glycan data comprising one or more glycan representations in the first platform-specific glycan format, the second platform-specific glycan format, or the universal glycan format; and execute the instructions to convert the input glycan data between: the first platform-specific glycan format and the universal glycan format; the second platform-specific glycan format and the universal glycan format; or the first platform-specific glycan format and the second platform-specific glycan format (e.g., via the universal glycan format).

[0058] In some embodiments, the processor is configured to execute the instructions to convert the input glycan data from the first platform-specific glycan format **410** to the second platform-specific glycan format **420** by applying: (1) the first rule set **415** to convert the input glycan data from the first platform-specific glycan format to the universal glycan format **400**; and (2) the second rule set **425** to convert the input glycan data from the universal glycan format to the second platform-specific glycan format. In some embodiments, the non-transitory memory further comprises instructions for obtaining the input glycan data from a first glycan database in the first platform-specific glycan format, and wherein execution of the instructions by the processor generates a second glycan database comprising glycan data in the second platform-specific glycan format. In some embodiments, the first platform-specific glycan format is compatible with a first glycomolecule search engine, e.g., a first glycopeptide search engine, but not compatible with a second glycomolecule search engine, e.g., a second glycopeptide search engine; and the second platform-specific glycan format is compatible with the second glycomolecule search engine, e.g., the second glycopeptide search engine.

[0059] In some embodiments, the plurality of format conversion rule sets further includes a third rule set **435** configured to convert glycan representations of a third platform-specific glycan format **430** to and from the universal glycan format **400**. In some embodiments, the glycan data management system includes a processor configured to: access an input glycan data comprising one or more glycan representations in the first platform-specific glycan format **410**, the second platform-specific glycan format **420**, the third platform-specific glycan format **430**, and/or the universal glycan format **400**; and execute the instructions to convert the input glycan data between: the first platform-specific glycan format and the universal glycan format; the second platform-specific glycan format and the universal glycan format; the third platform-specific glycan format and the universal glycan format; the first platform-specific glycan format and the second platform-specific

glycan format (e.g., via the universal glycan format); the first platform-specific glycan format and the third platform-specific glycan format (e.g., via the universal glycan format); or the second platform-specific glycan format and the third platform-specific glycan format (e.g., via the universal glycan format).

[0060] In some embodiments, the plurality of format conversion rule sets further includes an N^{th} rule set **445** configured to convert glycan representations of an N^{th} platform-specific glycan format **440** to and from the universal glycan format **400**, where N is an integer greater than 1. In some embodiments, the glycan data management system includes a processor configured to: access an input glycan data comprising one or more glycan representations in one or more of the first-to-N^{th} platform-specific glycan formats; and execute the instructions to convert the input glycan data between: any one of the first-to-N^{th} platform-specific glycan formats and the universal glycan format; or between any two of the first-to-N^{th} platform-specific glycan formats (e.g., via the universal glycan format). The glycan data management system can have instructions for converting among any suitable number of platform-specific glycan formats. In some embodiments, the number of platform-specific glycan formats is 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50, 60, 70, 80, 90, 100 or more, or a number in a range defined by any two of the preceding values.

[0061] Also provided are methods, e.g., computer-implemented methods, for converting a glycan data format. In general terms, the method involves converting a glycan data format from a first platform-specific glycan format to a second platform-specific glycan format via a universal glycan format. As shown in Fig. 4, the glycan data format can initially be converted from the first platform-specific glycan format **410** to the universal glycan format **400**, and then from the universal glycan format to the second platform-specific glycan format **420**. As discussed herein, the method can be expanded to provide conversion among the existing platform-specific glycan formats and a new platform-specific glycan format by adding a conversion rule set for converting the new platform-specific glycan format to the universal glycan format.

[0062] With reference to Fig. 5, a method **500** of converting a glycan data format is provided. The method can include receiving **510** first glycan data that includes one or more glycan representations in a first platform-specific glycan format. The method can further include converting **520** the first glycan data from the first platform-specific glycan format to a second platform-specific glycan format via a universal glycan format, where the second platform-specific glycan format, but not the first platform-specific glycan

format, is compatible with a first glycomolecule search engine. In some embodiments, the first glycomolecule search engine is a first glycopeptide search engine. In some embodiments, the first platform-specific glycan format is compatible with a second glycomolecule search engine, e.g., a second glycopeptide search engine. In some embodiments, the second platform-specific glycan format is not compatible with the second glycomolecule search engine, e.g., the second glycopeptide search engine.

[0063] In some embodiments, the method includes converting the first glycan data from the first platform-specific glycan format to a third platform-specific glycan format via the universal glycan format, where the third platform-specific glycan format, but not the first platform-specific glycan format, is compatible with a third glycomolecule search engine, e.g., a third glycopeptide search engine. In some embodiments, the method includes converting the first glycan data from the first platform-specific glycan format to an N^{th} platform-specific glycan format via the universal glycan format, where N is an integer greater than 1. A given glycan data can be converted to one or more of any suitable number of platform-specific glycan formats. In some embodiments, the number of platform-specific glycan formats is 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50, 60, 70, 80, 90, 100 or more, or a number in a range defined by any two of the preceding values.

[0064] In some embodiments, converting the first glycan data involves determining that the first glycan data includes the one or more glycan representations in the first platform-specific glycan format; and applying to the first glycan data: (1) a first rule set for converting the first platform-specific glycan format to the universal glycan format; and (2) a second rule set for converting the universal glycan data format to the second platform-specific glycan format. In some embodiments, the type of platform-specific glycan format of the first glycan data is provided by a user, e.g., as an input made by the user performing the method on a computer system. In some embodiments, the method includes detecting one or more features of the one or more glycan representations in the first glycan data and identifying the detected feature(s) as being distinguishing feature(s) of the first platform-specific glycan format. In some embodiments, determining that the first glycan data includes the one or more glycan representations in the first platform-specific glycan format is automated, e.g., does not require user input. In some embodiments, determining that the first glycan data includes the one or more glycan representations in the first platform-specific glycan format is done automatically by comparing the glycan representations with one or more known platform-specific glycan

formats (e.g., a known search engine-specific format, such as, but not limited to, PGLYCO3, BYONIC, and METAMORPHEUS). In some embodiments, determining that the first glycan data includes the one or more glycan representations in the first platform-specific glycan format involves the user selecting the first platform-specific glycan format, e.g., from a list of known platform-specific glycan formats.

[0065] Also provided is a method, e.g., computer-implemented method, of expanding the cross-compatibility of a glycan database, e.g., for use in a glycomolecule search engine. A glycan database may be compatible with a particular search engine but may not be compatible with a different search engine. As disclosed herein, glycan data management systems and methods of the present disclosure can provide cross-compatibility among different search engines for which the conversion rules for converting the platform-specific glycan format and the universal glycan format are known. When a new glycomolecule search engine using a new glycan database in a new glycan format is developed, the cross-compatibility of a glycan database with the new glycomolecule search engine can be implemented by adding a new rule set for converting the new glycan format to and from the universal glycan format. In some embodiments, the method includes providing to a glycan data management system of the present disclosure a glycan database that includes input glycan data having glycan representations in a first platform-specific glycan format; and adding to the plurality of format conversion rule sets a third rule set configured to convert glycan representations of a third platform-specific glycan format to and from the universal glycan format, wherein none of the other rule sets among the plurality of format conversion rule sets is configured to convert glycan representations of the third platform-specific glycan format to and from the universal glycan format, wherein the third platform-specific glycan format is compatible with a third glycomolecule search engine and wherein the first platform-specific glycan format is not compatible with the third glycomolecule search engine. In some embodiments, the method includes causing the processor to access the input glycan data of the glycan database and executing the instructions to convert the input glycan data to the third platform-specific glycan format via the universal glycan format. The cross-compatibility of glycan formats can be expanded for any number of new glycan formats by adding a single conversion rule set for each new glycan format to the system.

[0050] With reference to Fig. 23, a non-limiting embodiment of a method **2300** of identifying glycomolecules in a biological sample is provided. The method can include receiving **2310** a first search output for a biological sample, the first search output being

associated with a first glycomolecule search engine; and determining **2320** that the first search output identifies a first set of glycomolecules as being of a first type and a second set of glycomolecules as being of a second type. The first type may be different from the second type. The method can further include receiving **2330** a second search output for the biological sample, the second search output being associated with a second glycomolecule search engine; and determining **2340** that the second search output identifies a third set of glycomolecules as being of the first type and a fourth set of glycomolecules as being of the second type. The method can also include identifying **2350** among the first and second search outputs a conflicting subset of glycomolecules, where the conflicting subset of glycomolecules includes: (a) glycomolecules that are present in both the first set and the fourth set; or (b) glycomolecules that are present in both the second set and the third set. For example, in a conflicting subset, a sample data (e.g., a mass spectrum) for an analyte may be associated with the first type of glycomolecule in the first search output, and the sample data (e.g., the same mass spectrum) for the same analyte may be associated with the second type of glycomolecule in the second search output. This may be due, for example, to differences in the algorithms of the different search engines corresponding to the two search outputs. Similarly, for example, in a conflicting subset, a sample data for an analyte may be associated with the second type of glycomolecule in the first search output, and the sample data for the same analyte may be associated with the first type of glycomolecule in the second search output. In some embodiments, at least one or more of the glycomolecules in the conflicting subset of glycomolecules are present in both the first set and the fourth set, and at least one or more of some other glycomolecules in the conflicting subset of glycomolecules are present in both the second set and the third set. Figs. 6A and 6B and the associated description below illustrate an example of identifying a conflicting subset present in two search outputs.

[0051] In some embodiments, the conflicting subset of glycomolecules includes only glycomolecules that are present in both the first set and the fourth set. In some embodiments, the conflicting subset of glycomolecules includes only glycomolecules that are present in both the second set and the third set. In some embodiments, the first set of glycomolecules is mutually exclusive of the third set of glycomolecules. For example, a sample data for an analyte may be associated with the first type of glycomolecule in the first search output, and the sample data for the same analyte is not associated with the first type of glycomolecule in the second search output. In some embodiments, the second set of glycomolecules is mutually exclusive of the fourth set of glycomolecules. For example,

a sample data for an analyte may be associated with the second type of glycomolecule in the first search output, and the sample data for the same analyte is not associated with the second type of glycomolecule in the second search output.

[0052] Identifying the conflicting subsets can be done using any suitable option. In some embodiments, the correspondence between a glycomolecule identified in one search output with a glycomolecule identified in another search output involves retrieving an identifier (in mass spectrometry, for example, this identifier may sometimes be referred to as a "scan number") for the glycomolecule identified in one search output, and using the identifier to find the corresponding glycomolecule identified in the other search output. In some embodiments, the first search output includes: a first set of identifiers (e.g., scan numbers) associated with each glycomolecule of the first set of glycomolecules; and a second set of identifiers (e.g., scan numbers) associated with each glycomolecule of the second set of glycomolecules, and the second search output includes: a third set of identifiers (e.g., scan numbers) associated with each glycomolecule of the third set of glycomolecules; a fourth set of identifiers (e.g., scan numbers) associated with each glycomolecule of the fourth set of glycomolecules. In some embodiments, identifying the conflicting subset of glycomolecules includes: (a) determining an identifier in the first set of identifiers that is the same as an identifier in the fourth set of identifiers; or (b) determining an identifier in the second set of identifiers that is the same as an identifier in the third set of identifiers. For example, in a conflicting subset, a sample data for an analyte having a specific identifier may be associated with a specific identifier and with the first type of glycomolecule in the first search output, and the sample data for the same analyte having the same specific identifier may be associated with the specific identifier and with the second type of glycomolecule in the second search output. Similarly, for example, in a conflicting subset, a sample data for an analyte may be associated with a specific identifier and with the second type of glycomolecule in the first search output, and the sample data for the same analyte may be associated with the same specific identifier and with the first type of glycomolecule in the second search output. In some embodiments, the identifier is a scan number associated with mass spectrometry data for analytes of a sample.

[0053] Search outputs from any suitable number of glycomolecule search engines can be analyzed. In some embodiments, the method further includes receiving an N^{th} search output for the biological sample, the N^{th} search output being associated with an N^{th} glycomolecule search engine; and determining that the N^{th} search output identifies a $(2N-1)^{th}$ set of glycomolecules as being of a first type and a $2N^{th}$ set of glycomolecules as

being of a second type, where N is an integer greater than 0 (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 50 or more, or a number in a range defined by any two of the preceding values). The first type may be different from the second type. In some embodiments, the method then includes identifying, among any two or more (e.g., 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 50 or more, or a number in a range defined by any two of the preceding values) of the N search outputs, a conflicting subset of glycomolecules. For example, in the case where there are three search outputs, a conflicting subset may be identified among sets corresponding to two of the three search outputs, or among sets corresponding to all three search outputs. In some embodiments, the conflicting subset includes glycomolecules that are present in one or more $(2k\text{-}1)^{th}$ set of the 2N sets of glycomolecules, and one or more $2m^{th}$ set of the 2N sets of glycomolecules, where k is an integer greater than 0, $(2k\text{-}1)$ is less than or equal to N, m is an integer greater than 0, $2m$ is less than or equal to N, and $k \neq m$. In some embodiments, at most 95%, e.g., at most 90%, at most 80%, at most 70%, at most 60%, at most 50%, or at most a percentage in a range defined by any two of the preceding values, of the search outputs identify a glycomolecule of the conflicting set as being of the same type (e.g., the first type), and one or more of the other search outputs identify the glycomolecule of the conflicting set as being of a different type (e.g., the second type).

[0054] With reference to Fig. 24A, a non-limiting example of a method **2400** of identifying glycomolecules in a biological sample is provided. The method can include receiving **2410** a first search output for a biological sample, the first search output being associated with a first glycomolecule search engine; and receiving **2420** a second search output for the biological sample, the second search output being associated with a second glycomolecule search engine. The method can further include determining **2430** that the first and second search outputs identify a conflicting subset of glycomolecules, where the conflicting subset of glycomolecules includes one or more glycomolecules determined as being both of a first type in the first search output and of a second type in the second search output. The first type may be different from the second type. For example, in a conflicting subset, a sample data for an analyte may be associated with the first type of glycomolecule in the first search output, and the sample data for the same analyte may be associated with the second type of glycomolecule in the second search output. In some embodiments, the one or more glycomolecules of the conflicting subset of glycomolecules are determined as not being of the second type in the first search output. In some embodiments, the one or

more glycomolecules of the conflicting subset of glycomolecules are determined as not being of the first type in the second search output.

[0055] As discussed above, identifying the conflicting subsets can be done using any suitable option, e.g., by using a suitable identifier (e.g., scan number) associated with an analyte identified in multiple search outputs. In some embodiments, the first search output includes a set of identifiers (e.g., scan numbers) associated with each glycomolecule of a first set of glycomolecules, and the second search output includes a second set of identifiers (e.g., scan numbers) associated with each glycomolecule of a second set of glycomolecules. In some embodiments, determining that the first and second search outputs identify the conflicting subset of glycomolecules involves: identifying a first identifier in the first set of identifiers associated with a glycomolecule of the first type in the first set of glycomolecules; and identifying a second identifier in the second set of identifiers associated with a glycomolecule of the second type in the second set of glycomolecules, where the first and second identifiers are the same. For example, in a conflicting subset, a sample data for an analyte may be associated with a specific identifier and with the first type of glycomolecule in the first search output, and the sample data for the same analyte may be associated with the specific identifier and with the second type of glycomolecule in the second search output. In some embodiments, the identifier is a scan number.

[0056] Search outputs from any suitable number of glycomolecule search engines can be analyzed. In some embodiments, the method further includes receiving an N^{th} search output for the biological sample, the N^{th} search output being associated with an N^{th} glycomolecule search engine; and receiving an $(N+1)^{th}$ search output for the biological sample, the $(N+1)^{th}$ search output being associated with an $(N+1)^{th}$ glycomolecule search engine, where N is an integer greater than 0, e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 50 or more, or a number in a range defined by any two of the preceding values. In some embodiments, the method then includes identifying among any two or more, e.g., 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 50 or more, or a number in a range defined by any two of the preceding values of the N search outputs a conflicting subset of glycomolecules. In some embodiments, the conflicting subset of glycomolecules includes one or more glycomolecules determined in at most 95%, e.g., at most 90%, at most 80%, at most 70%, at most 60%, at most 50%, or at most a percentage in a range defined by any two of the preceding values, of the search outputs as being of a first type and of a second type in one or more of the other search outputs.

[0057] The search output obtained from the search engines can include any suitable information related to or associated with each identified glycomolecule. In general, the search output includes information about a glycan identified in the sample, e.g., glycan data, based on the glycomolecule sample data, e.g., mass spectrometry data, of the biological sample. In some embodiments, the glycan data includes glycan compositional information. In some embodiments, the glycan data includes glycan structural information. In some embodiments, the glycan data does not include glycan structural information. In some embodiments, the glycan data includes glycan compositional information only. In some embodiments, the glycan data includes the amount of hexose, N-acetylhexosamine, fucose, sialic acid (e.g., N-acetylneuraminic acid, N-glycolylneuraminic acid), and/or any other suitable sugar associated with a glycomolecule.

[0058] In some embodiments, the search output includes, without limitation, one or more of: glycan data, an identifier (in the field of mass spectrometry for example, the identifier is sometimes referred to as a "scan number"), a peptide sequence, a protein ID, a retention time, an m/z value, a charge state (also called "charge number"), and a confidence score. The identifier can be used to identify an analyte for which sample data, e.g., mass spectrometry data, was obtained, and can serve as a label to identify the same analyte in search outputs obtained from different glycomolecule search engines, e.g., glycopeptide search engines, such as, without limitation, PGLYCO, BYONIC, METAMORPHEUS, STRUCGP, etc., that are run on the same mass spectrometry data. In some embodiments, the glycan data provides compositional and/or structural information for glycans determined as being likely present in a sample (e.g., glycans of glycopeptides, glycolipids, glycoRNA, glycoDNA) based on the sample data, e.g., mass spectrometry data. In embodiments where a glycopeptide search is performed, the search output may include glycan data and the peptide sequence of the polypeptide to which a glycan is conjugated, based on the sample data. In some embodiments, the retention time, the m/z value, and/or the charge state are provided for the analyte (which may be, for example, a glycopeptide) that is identified based on the sample data.

[0059] The search output can be in any suitable format that provides confidence scores and that can be used to determine the overlap. In some embodiments, the search output is in a search engine-specific format. In some embodiments, the search output is converted to a common format, e.g., to facilitate the analysis. In some embodiments, the common format is a universal glycomolecular data format. In some embodiments, the first search output is in a first format, the second search output is in a second format, and the

method further includes, before determining the glycomolecule profile, converting one or more of the first and second search outputs to a common format, where the common format comprises the first format, the second format, or a universal glycomolecular data format.

[0060] In some embodiments, the confidence score provides an indication of how well the data for a particular analyte matches the glycomolecule identified by the search engine. In other words, the confidence score may be an indication of a level of confidence the search engine's algorithm has that the correct glycomolecule was identified. In some embodiments, the confidence score in the search output is specific to the search engine. In some embodiments, the confidence score provides an indication of how well the data for a particular analyte matches the entire glycomolecule (e.g., the glycan and the non-glycan component of the glycomolecule) identified by the search engine. In some embodiments, the confidence score provides an indication of how well the data for a particular analyte matches the non-glycan component of the glycomolecule identified by the search engine. The confidence score may be determined by one or more algorithms of a respective search engine. In some embodiments, where the glycomolecule is a glycopeptide, the confidence score provides an indication of how well the data for a particular analyte matches the peptide component of the glycopeptide identified by the search engine. Due to differences in search engine algorithms and formats, different search engines may return different confidence scores for the same analyte such that search output confidence scores cannot be easily matched or compared among search outputs from different search engines. This is discussed in further detail below.

[0061] In some embodiments, the method includes generating a notification comprising a listing of the conflicting subset of glycomolecules. The notification can be generated using any suitable option. In some embodiments, the notification may be displayed on a screen, a computer monitor, a touch screen, a smartphone, or any other suitable device; is printed on paper; is read out through a speaker, etc. The notification can provide any suitable information associated with the listed conflicting subset of glycomolecules. In some embodiments, the notification includes, without limitation, one or more of: glycan data, an identifier, a peptide sequence, a precursor mass, a peptide sequence with other post-translational modifications (phosphorylation, acetylation, oxidation, carbamidomethylation, and other modifications), a protein ID, a retention time, an m/z value, a charge state, a confidence score (alternatively, a peptide confidence score and/or a glycan confidence score and/or an overall confidence score), and search engine identity associated with each of the identified glycomolecules in the listed conflicting

subset of glycomolecules. The notification can provide the listing of the conflicting subset of glycomolecules in any suitable format. In some embodiments, the listing of the conflicting subset of glycomolecules is provided in the notification in a search engine-specific format. In some embodiments, the listing of the conflicting subset of glycomolecules is provided in the notification in a common format, such as a universal glycomolecular data format. In some embodiments, the method includes providing a link to the subset of the sample data underlying the conflicting subset of glycomolecule that allows a user to pull out the relevant data for further review (e.g., further review of the mass spectrometry data), as disclosed herein. In some embodiments, the method includes providing a selectable option for the use to provide confirmation that the glycomolecule is of the first type, the second type, or neither type (which can be a third option, or may be an undetermined type).

[0062] In some embodiments, the method includes options for resolving the conflicting identification of glycomolecules between the different glycomolecule search engines. In some embodiments, from the data underlying the search results, e.g., the glycomolecule sample data or mass spectrometry data used as an input to the glycomolecule search engines, a subset that corresponds to the conflicting subset of glycomolecules is extracted and one or more additional analyses are performed on the subset of data. In some embodiments, the one or more additional analyses resolves the conflicting identification of glycomolecules and confirms their identity. In some embodiments, the method includes selecting a subset of data corresponding to the conflicting subset of glycomolecules; and performing an additional analysis on the selected data to confirm an identity of the one or more glycomolecules of the conflicting subset of glycomolecules. In some embodiments the additional analysis confirms that a glycomolecule of the conflicting subset of glycomolecules is of the first type or the second type. In some embodiments the additional analysis confirms that a glycomolecule of the conflicting subset of glycomolecules is neither the first type nor the second type. In some embodiments the additional analysis confirms that a glycomolecule of the conflicting subset of glycomolecules is of a third type that is different from the first or the second type.

[0063] In some embodiments, resolving the conflicting subset of glycomolecules involves an analysis of confidence scores associate with glycomolecules of the conflicting subset. In some embodiments, the first search output includes a first set of confidence scores associated with the first or the second set of glycomolecules, and the second search output includes a second set of confidence scores associated with the third

or the fourth set of glycomolecules. As noted above, the confidence score can be a search-engine specific confidence score. For example, the confidence score in the first search output may be a number on a scale of 0 to 100, while the confidence score in the second search output may be a number on a scale of 0 to 1000. In some embodiments, correspondence between confidence scores from different glycomolecule search engines is determined empirically. In some embodiments, the confidence scores may first be normalized before further analysis, e.g., before comparison of confidence scores. For example, in the case of the confidence score in the first search output on a scale of 0 to 100 and the confidence score in the second search output on a scale of 0 to 1000, the scores may be normalized so that they are both on a scale of 0 to 100 or on a scale of 0 to 1000. In some embodiments, a weighting function can be applied to the confidence scores before further analysis. In some embodiments, the weighting function depends on the glycomolecule search engine from which the search output is obtained. In some embodiments, the weighting function depends on the type or source of the sample from which the sample data underlying the search output was obtained and the glycomolecule search engine used to generate the search output. For example, for a particular search engine known to be less reliable for a particular sample type, e.g., stool, compared to another sample type, e.g., serum, the weighting function may suppress or reduce the confidence scores in the search output from that particular search engine for sample data obtained from a stool sample compared to a serum sample. In some embodiments, a weighting function for a glycomolecule search engine is determined empirically.

[0064] In some embodiments, a confidence score threshold score can be used to set a minimum confidence score required to be considered a reliable identification of the glycomolecule by the search engine. In some embodiments, each confidence score of the first set of confidence scores is equal to or greater than a first confidence score threshold, and each confidence score of the second set of confidence scores is equal to or greater than a second confidence score threshold. In some embodiments, the confidence score threshold is specific to each search engine. In some embodiments, the confidence score thresholds are scaled to the total range of each search-engine specific confidence scores. For example, in some embodiments a confidence score threshold of 20 for a search-engine specific confidence score that has a total range of 0 to 100 may be comparable to a confidence score threshold of 200 for another search-engine specific confidence score that has a total range of 0 to 1000.

[0065] In some embodiments, the confidence scores are compared to resolve the conflicting subset of glycomolecules. In some embodiments, the method includes resolving the conflicting subset of glycomolecules based on a comparison of confidence scores of the first and second sets of confidence scores. For example, for a glycomolecule identified as being of a first type in a first search output and of a second type in a second search output, the confidence score associated with the glycomolecule in the first search output can be compared with the confidence score associated with the glycomolecule in the second search output. In some embodiments, the first and second sets of confidence scores are normalized, e.g., to facilitate the comparison. In some embodiments, resolving the conflicting subset of glycomolecules comprises, for each glycomolecule identified as being present in both the first set and the fourth set, or in both the second set and the third set, determining that the glycomolecule is of the type identified by the search output comprising the higher confidence score associated with the glycomolecule. In some embodiments, the confidence score is higher (e.g., on a normalized scale) by at least 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, 150%, 200%, 250%, 300%, 400%, 500%, 750%, 1000% or more, or is higher by a percentage in a range defined by any two of the preceding values.

[0066] In some embodiments, the additional analysis to resolve the conflicting subset of glycomolecules involves review of the sample data underlying the search output by a researcher or user. For example, a user may visually inspect the mass spectra associated with the analyte that gave rise to the conflicting identification by different glycomolecule search engines, and confirm that the glycomolecule is of the first type, second type. In some embodiments, the glycomolecule is confirmed to be neither the first nor second type. In some embodiments, the glycomolecule is confirmed to be a third type that is different from the first or second type.

[0067] In some embodiments, the additional analysis to resolve the conflicting subset of glycomolecules involves running a search with the same sample data, or the subset thereof associated with the conflicting subset of glycomolecules, on a different glycomolecule search engine (e.g., a glycomolecule search engine different from the first and second glycomolecule search engines). In some embodiments, if the third glycomolecule search engine's output identifies the glycomolecule as being the first type, the glycomolecule is confirmed as being the first type. In some embodiments, if the third glycomolecule search engine's output identifies the glycomolecule as being the second type, the glycomolecule is confirmed as being the second type. Any number of additional

glycomolecule search engine can be relied upon to further analyze the sample data. In some embodiments, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50 or more, or a number in a range defined by any two of the preceding values, of different glycomolecule search engines can be used to further analyze the sample data and resolve the conflicting subset of glycomolecules.

[0068] The conflicting subset of glycomolecules can include any suitable types of glycomolecules that can be identified as a first type by a glycomolecule search engine and can be identified as a second type by another glycomolecule search engine, based on a suitable sample data (e.g., glycomolecule mass spectrometry data). Without being bound by theory, the features of the sample data that a search engine relies on to predict or determine the properties or identities of glycomolecules in a sample can, in some cases, lead to mis-identification of the type of glycomolecule. As discussed above, different search engines will typically rely on different algorithms, and may rely on different features of the data and/or weigh features differently to perform the search. Conflicting subsets can then be flagged for further validation or action, as described further below.

[0069] In some embodiments, the methods and systems disclosed herein may consider particular types of glycomolecule pairs more closely than others to identify conflicting subsets that are more prone to mis-identification. For example, the methods and systems may weight conflicting subsets of particular types more than other in flagging for further validation or action. Different types of glycomolecules that are characterized by features in the sample data that are similar to each other are more likely to be misidentified, and may this be considered more closely. In some embodiments, glycomolecules of the first type and glycomolecules of the second type are similar in one or more features of the sample data, including, without limitation, mass, retention time, m/z value, charge state, or any combination thereof. For example, glycomolecules of the first type and glycomolecules of the second type have a difference in mass of 0 to 1 daltons. In some embodiments, glycomolecules of the first type and glycomolecules of the second type have a difference in mass of 0 to 2 daltons. In some embodiments, glycomolecules of the first type and glycomolecules of the second type have a difference in mass of 0 to 3 daltons. In some embodiments, glycomolecules of the first type and glycomolecules of the second type have a difference in mass of 1, 2, 3, 4, 5, 6, 7, or 8 daltons, or a number in a range defined by any two of the preceding values. In some embodiments, glycomolecules of the first type and glycomolecules of the second type have different types of glycans associated with (e.g., conjugated to) the glycomolecule. In some cases, these glycomolecules may have similar

masses such that they may be easily confused. For example, in an MRM-MS experiment, a glycopeptide having a glycan of the first type may have a mass that would be similar to the mass of a glycopeptide having a glycan of the second type.

[0070] Figs. 24B-24D examples of different pairs of glycans with similar masses. In some embodiments, the first type is a sialic acid-containing glycomolecule (e.g., a sialic acid-containing glycopeptide) and the second type is a fucosylated glycomolecule (e.g., a fucosylated glycopeptide). In some embodiments, the sialic acid is N-acetylneuraminic acid. An example structure of such a pairing is illustrated in Fig. 24B, where the glycans are identical except that the first glycan includes two fucoses and the second glycan includes one N-acetylneuraminic acid. Although these glycans are different, they have very similar masses (resulting in m/z of 1955.6972 and 1954.6768) because the mass of two fucoses is approximately the same as the mass of one N-acetylneuraminic acid. Due to the similar masses, one may be confused for the other. This confusion can have significant impact in disease diagnostics for example, where a glycopeptide having sialic acid may be a biomarker for a disease while a glycopeptide having fucose may not be a biomarker for that disease. Other pairs of glycans that have similar masses (and may thus cause confusion) are illustrated in Figs. 24C and 24D. As another example, 5 hexose has a similar mass as 4 N-acetylhexosamine, such that a glycan such as N6H7 may be confused with N2H12. As another example, 2 N-acetylhexosamine + 1 fucose + 2 N-acetylneuraminic acid has a similar mass as 7 hexose, such that a glycan such as N2H12F0S0G0 may be confused with N4H5F1S2. As another example, 2 N-acetylhexosamine + 2 hexose have similar mass as 1 fucose + 2 N-acetylneuraminic acid. Although this disclosure mentions misidentification between particular glycomolecule pairs, the disclosure contemplates other suitable glycomolecule pairs that may be misidentified. Similarly, although the disclosure focuses on pairs of easily misidentified glycomolecules, the disclosure contemplates any suitable grouping of misidentified glycomolecules (e.g., three types of glycomolecules that are easily mistaken for each other).

[0071] Also provided herein is a method of determining a glycomolecule profile of a biological sample, e.g., based at least in part on an analysis of confidence scores. With reference to Fig. 25, a non-limiting embodiment of the method **2500** is provided. The method can include receiving **2510** a first search output for a biological sample, the first search output being associated with a first glycomolecule search engine. The first search output can include a first set of glycomolecules associated with a first set of confidence scores. The method can also include receiving **2520** a second search output for the

biological sample, the second search output being associated with a second glycomolecule search engine. The second search output can include a second set of glycomolecules associated with a second set of confidence scores. The method can further include determining **2530** a glycomolecule profile of the biological sample based on (1) an overlap between the first and second search outputs, and (2) the first and second set of confidence scores. Without being bound by theory, agreement in the identification of glycomolecules between two or more glycomolecule search engine run on the same sample data can indicate that the identification is reliable. This may increase the confidence level of a user has on the reliability of the search result relative to when search results from only one glycomolecule search engine are used. The number of search outputs received is not particularly limited. In some embodiments, the method includes receiving 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 40, 50, or more, or a number in a range defined by any two of the preceding values, search outputs, each associated with a different glycomolecule search engine. In some embodiments, the overlap can be determined for any suitable number of search outputs. In some embodiments, the overlap is between any two or more of three search outputs. For example, the overlap can be between the first and second search outputs; between the second and third search outputs; between the first and third search outputs; or between the first, second, and third search outputs. In some embodiments, determining the glycomolecule profile is based on an overlap between any 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 40, 50, or more, or a number in a range defined by any two of the preceding values, of the search outputs, or all of the search outputs, and the corresponding sets of confidence scores. As noted above, the search output obtained from the search engines can include any suitable information related to or associated with each identified glycomolecule, and can be in any suitable format.

[0072] As discussed above, the confidence score can provide an indication of how well the data for a particular analyte matches the glycomolecule identified by the search engine. In some embodiments, a weighting function can be applied to the confidence scores. In some embodiments, the weighting function depends on the glycomolecule search engine from which the search output is obtained. In some embodiments, the weighting function depends on the type or source of the sample from which the sample data underlying the search output was obtained and the glycomolecule search engine used to generate the search output. In some embodiments, a weighting function for a glycomolecule search engine is determined empirically.

[0073] In some embodiments, the overlap between the search outputs includes glycomolecules that are identified as being the same among the search outputs. In general terms, an overlap is found between different search outputs when there is agreement in identification of one or more parameters of the search outputs for the same analyte, e.g., for identified glycomolecules having the same identifier. In some embodiments, the same glycomolecules are identified when the identified glycans are the same. In some embodiments, the same glycomolecules are identified when the identified peptide sequences are the same. In some embodiments, the same glycomolecules are identified when the identified glycans and peptide sequences are both the same. In some embodiments, the overlap includes one or more of peptide sequences, protein IDs, retention times, m/z values, or charge states.

[0074] In some embodiments, a threshold score for the confidence score for each search output can be used to focus the analysis on those glycomolecules identified by the respective search engines with a desired degree of confidence. In some embodiments, the overlap includes a set of glycomolecules in both the first and second search outputs. In some embodiments, each glycomolecule in the first set of glycomolecules is associated with a first confidence score of the first set of confidence scores that is equal to or greater than a first confidence score threshold. In some embodiments, each glycomolecule in the second set of glycomolecules is associated with a second confidence score of the second set of confidence scores that is equal to or greater than a second confidence score threshold. The confidence score thresholds (e.g., the first confidence score threshold, second confidence score threshold) can be any suitable value. In some embodiments, the confidence score thresholds (e.g., the first confidence score threshold, second confidence score threshold) are specific to each search engine. In some embodiments, the confidence score thresholds are scaled to the total range of each search-engine specific confidence scores. For example, in some embodiments a confidence score threshold of 20 for a search-engine specific confidence score that has a total range of 0 to 100 may be comparable to a confidence score threshold of 200 for another search-engine specific confidence score that has a total range of 0 to 1000.

[0075] In some embodiments, the method includes adjusting confidence score thresholds of the first and/or second search outputs to change the proportion or percentage of overlap between the first and second sets of glycomolecules relative to the total number of glycomolecules identified with a confidence score above the threshold. In some embodiments, confidence score thresholds of the first and/or second search outputs is

increased to increase the proportion or percentage of overlap between the first and second sets of glycomolecules relative to the total number of glycomolecules identified with a confidence score above a confidence score threshold. In some embodiments, confidence score thresholds of the first and/or second search outputs is decreased to increase the proportion or percentage of overlap between the first and second sets of glycomolecules relative to the total number of glycomolecules identified with a confidence score above a confidence score threshold. The method may include performing any suitable combination of increasing or decreasing the confidence score thresholds of the first and/or second search outputs to increase the proportion or percentage of overlap. In some embodiments, the method includes setting the value of the first confidence score threshold and the value of the second confidence score threshold to optimize a percentage overlap between the search outputs among all the first and second glycomolecules having, respectively, first confidence scores equal to or greater than the first confidence score threshold and second confidence scores equal to or greater than the second confidence score threshold. In some embodiments, the method may include optimizing an overlap between two or more search outputs. In some embodiments, the optimized overlap is at least 50%, at least 60%, at least 70%, at least 75%, at least 80%, at least 90%, or at least 95%, or a percentage in a range defined by any two of the preceding values. In some embodiments, the confidence score threshold is set to include glycomolecules identified with confidence score in the top 1%, 2%, 3%, 4%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95% or more, or a percentage in a range defined by any two of the preceding values, in the search output. Figs. 29A and 29B and the associated description below illustrate an example of adjusting threshold scores to optimize overlaps.

[0076] As discussed above, in some embodiments, the search output includes an identifier (e.g., a scan number), that can be used to trace back a search result regarding a particular glycomolecule identified in one search output to the subset of the sample data underlying the search result, and to find the corresponding search result in a different search output (from a different glycomolecule search engine run on the same sample data). In some embodiments, the first search output includes a first set of identifiers, and the second search output includes a second set of identifiers. In some embodiments, determining the glycomolecule profile further includes: determining the first set of identifiers associated with the first set of glycomolecules; and determining the second set of identifiers associated with the second set of glycomolecules.

[0077] In some embodiments, the method further includes generating a consensus list of glycomolecules based on the overlap. The consensus list of glycomolecules can be generated using any suitable option. In some embodiments, a consensus list of glycomolecules is generated by determining an overlap between search outputs and/or determining differences (e.g., non-overlapping portions) between search outputs. As discussed above, in some embodiments, the consensus list can include all glycomolecules identified as being in the first set and the second set. In some embodiments, generating the consensus list includes applying a weighting factor to either one or both of the first and second sets of confidence scores. In some embodiments, the weight of the contribution of each glycomolecule search engine to the consensus list of glycomolecules differs between different search engines. In some embodiments, determining the glycomolecule profile includes applying a weighting factor to the search output according to the glycomolecule search engine associated therewith.

[0078] In some embodiments, the consensus list of glycomolecules includes a confidence score for each identified glycomolecule. In some embodiments, the confidence score in the consensus list is a search-engine specific confidence score. In some embodiments, the confidence score in the consensus list is a consensus confidence score derived from the search engine-specific confidence scores of two or more search engines. For example, in the case of two search outputs from two search engines, for each analyte (corresponding to the same identifier in both search outputs), a consensus confidence score may be determined by applying a function to the two search engine-specific confidence scores of the two search outputs. For example, the consensus confidence score may be obtained by computing an average of the search engine-specific confidence scores for an analyte. As another example, the consensus confidence score may be obtained by computing a median of search engine-specific confidence scores an analyte. As another example, the consensus confidence score may be obtained by computing a weighted function (e.g., a weighted average) of search engine-specific confidence scores for an analyte. Computing a weighted function may include applying one or more different weighting factors to different search engines. In some embodiments, the confidence scores may first be normalized before applying a desired function (e.g., an average function, a median function, a weighted average function) to derive the consensus confidence score. For example, in the case of a first search engine-specific confidence score on a scale of 0 to 100 and a second search engine-specific confidence score on a scale of 0 to 1000, the

scores may be normalized so that they are both on a scale of 0 to 100 or on a scale of 0 to 1000.

[0079] The sample or biological sample in any method or system of the present disclosure can be any suitable biological source of glycomolecules (e.g., glycopeptide, glycoprotein, glycolipid, glycoRNA, glycoDNA). In some embodiments, the biological sample is a urine, stool, blood, or serum sample. In some embodiments, the sample is from human urine, human stool, human blood, or human serum. In some embodiments, the sample is from yeast, a plant, a non-human mammal, or a human. In some embodiments, a method of the present disclosure includes generating the glycomolecule sample data or mass spectrometry data from the biological sample.

[0080] Any method of the present disclosure can be performed on a suitable system configured to perform the various steps. Any method of the present disclosure can be implemented on a suitable computer system. Provided herein is an electronic system that includes: a processor; and non-transitory memory comprising instructions, which when executed by the processor causes the processor to perform any method of the present disclosure.

Glycan formats

[0066] Glycan data, e.g., input glycan data, for use in the systems and methods of the present disclosure can be any suitable data providing compositional and/or structural information of glycans, e.g., in a glycan database. With reference to Fig. 4, the glycan format (e.g., the universal glycan format **400**, any one or more of the platform-specific glycan formats **410**, **420**, **430**, **440**) can be any suitable representation of glycan data. The glycan format (e.g., universal glycan format **400**, any one or more of the platform-specific glycan formats **410**, **420**, **430**, **440**) can include a representation of any suitable aspect of a glycan in the glycan data. The glycan format (e.g., universal glycan format **400**, any one or more of the platform-specific glycan formats **410**, **420**, **430**, **440**) can include glycan compositional information. In some embodiments, the universal glycan format includes compositional information only. In some embodiments, the universal glycan format includes glycan structural information. In some embodiments, the universal glycan format includes glycan compositional information and glycan structural information. In some embodiments, at least one of the platform-specific glycan formats includes compositional information only. In some embodiments, at least one of the platform-specific glycan formats includes glycan structural information. In some embodiments, at least one of the

platform-specific glycan formats includes glycan compositional information and glycan structural information (e.g., the format compatible with the PGLYCO3 search engine).

[0067] The universal glycan format **400** can include any suitable representation for different monosaccharides that are or may be present in a glycan. In some embodiments, the universal glycan format includes a single-letter representation for different monosaccharides. In some embodiments, the universal glycan format uses 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more characters to represent each of the different monosaccharides. In some embodiments, the universal glycan format represents hexose as "H". In some embodiments, the universal glycan format represents N-acetylhexosamine as "N". In some embodiments, the universal glycan format represents fucose as "F". In some embodiments, the universal glycan format represents N-acetylneuraminic acid as "A". In some embodiments, the universal glycan format represents N-glycolylneuraminic acid as "G". In some embodiments, each representation of a monosaccharide in the universal glycan format is associated with the amount of the monosaccharide in a glycan.

[0068] In some embodiments, the glycan format (e.g., the universal glycan format **400**, or any one or more of the platform-specific glycan formats **410**, **420**, **430**, **440**) provides the amount of a monosaccharide that is present in a glycan. In some embodiments, the glycan format (e.g., the universal glycan format **400**, any one or more of the platform-specific glycan formats **410**, **420**, **430**, **440**) lists the number of each monosaccharide that is present in the glycan. In some embodiments, the universal glycan format **400** lists the amount of a monosaccharide that is not present in the glycan, e.g., as "0". Inclusion of monosaccharides that are not present in the glycan representation is advantageous because it ensures uniformity in the format of the representation. That is, in the universal glycan format, glycan representations for each glycan include information about all the same sugars of interest (whether present or absent). This is in contrast to conventional formats, where glycan representations omit the monosaccharides that are absent. This uniformity of the universal glycan format is readily evident in Figs. 9 and 10, which show a number of glycan representations in conventional platform-specific formats and their counterparts in the universal glycan format. The uniformity allows for better readability, and also simplifies comparison and analysis of multiple glycomolecules. In some embodiments, the universal glycan format includes the amount of fucose and/or sialic acid to represent a glycan, even if there is no fucose and/or sialic acid present in the glycan. In some embodiments, the universal glycan format includes the amount of N-acetylneuraminic acid and/or N-glycolylneuraminic acid to represent a glycan, even if there is no N-

acetylneuraminic acid and/or N-glycolylneuraminic acid present in the glycan. In some embodiments, the universal glycan format includes a first number *l* corresponding to an amount of hexose in the glycan data; a second number *m* corresponding to an amount of N-acetylhexosamine in the glycan data; a third number *n* corresponding to an amount of fucose in the glycan data; and a fourth number *o* corresponding to an amount of N-acetylneuraminic acid in the glycan data, where *l*, *m*, *n*, and *o* are each an integer 0 or greater. In some embodiments, the universal glycan format further comprises a fifth number *p* corresponding to an amount of N-glycolylneuraminic acid, wherein *p* is an integer 0 or greater.

[0069] In some embodiments, a glycan format (e.g., the universal glycan format **400**, any one or more of the platform-specific glycan formats **410**, **420**, **430**, **440**) may provide glycan structural information. For example, referencing Fig. 9, the PGLYCO (Structure) format uses the sequence of letters to represent the relative position of associated monosaccharides and further uses parentheses to denote the branching structure of the glycan. In some embodiments, the glycan format includes, for a given glycan, glycan structural information. In some embodiments, at least one of the platform-specific glycan formats (e.g., **410**, **420**, **430**, **440**) includes glycan structural information. In some embodiments, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50, 60, 70, 80, 90, 100 or more, or a number in a range defined by any two of the preceding values, or the platform-specific glycan formats includes glycan structural information. In some embodiments, none of the platform-specific glycan formats includes glycan structural information. In some embodiments, the universal glycan format **400** includes glycan structural information. In some embodiments, the universal glycan format **400** does not include glycan structural information.

[0070] In some embodiments, the glycan structural information is, for a given glycan, a single glycan structural arrangement. In some embodiments, the glycan structural information is, for a given glycan, the most likely glycan structural arrangement. In some embodiments, the glycan structural information is, for a given glycan, a set of all possible glycan structural arrangements. In some embodiments, the glycan structural information is, for a given glycan, a set of all biologically consistent glycan structural arrangements (e.g., consistent with the type of structural arrangements known to be found in glycans from the same biological source).

[0071] The platform-specific glycan format (e.g., **410**, **420**, **430**, **440**) can be suitable for representing glycan data as an input to a computer-implemented glycomolecule

analysis routine, such as, but not limited to, a glycomolecule search engine. In some embodiments, the platform-specific glycan format (e.g., **410**, **420**, **430**, **440**) is a glycan format used by or compatible with a glycomolecule search engine. In general, a glycan database can be an input to a glycomolecule search engine when glycan data of the glycan database are represented in the glycan format used by or compatible with the glycomolecule search engine. In general, a glycan database cannot be an input to a glycomolecule search engine when glycan data of the glycan database are represented in a glycan format that is not used by or is incompatible with the glycomolecule search engine. In some embodiments, the first platform-specific glycan format **410** is compatible with (e.g., a glycan database in the first platform-specific glycan format can be an input **460** to) a glycomolecule search engine **450**. In some embodiments, the second platform-specific glycan format **420** is not compatible with (e.g., a glycan database in the second platform-specific glycan format cannot be an input **470** to) the glycomolecule search engine **450** with which the first platform-specific glycan format is compatible. In some embodiments, only the first platform-specific glycan format **410** is compatible with the glycomolecule search engine **450**. In some embodiments, the second platform-specific glycan format **420** is compatible with (e.g., a glycan database in the second platform-specific glycan format can be an input to) a glycomolecule search engine with which the first platform-specific glycan format is not compatible.

[0072] In some embodiments, the platform-specific glycan format is a glycan format used by or compatible with a particular glycomolecule search engine or a particular subset of glycomolecule search engines that is available for public and/or commercial use. In some embodiments, the glycomolecule search engine is a glycopeptide search engine. In some embodiments, the platform-specific glycan format is a glycan format used by or compatible with a glycopeptide search engine, such as, but not limited to BYONIC, PGLYCO (e.g., PGLYCO3), or METAMORPHEUS.

[0073] In some embodiments, the universal glycan format is different from at least one of the platform-specific glycan formats. In some embodiments, the universal glycan format is different from any of the platform-specific glycan formats. In some embodiments, the universal glycan format is the same as one of the platform-specific glycan formats.

Format conversion rule sets

[0074] The format conversion rule sets (e.g., **415**, **425**, **435**, **445**) can include any suitable collection of conversion rules for converting representation of glycan data from one format to another. In some embodiments, the first rule set **415** includes conversion rules for converting the first platform-specific glycan format **410** into the universal glycan format **400**. In some embodiments, the first rule set includes conversion rules for converting the universal glycan format **400** into the first platform-specific glycan format **410**. In some embodiments, the second rule set **425** includes conversion rules for converting the second platform-specific glycan format **420** into the universal glycan format **400**. In some embodiments, the second rule set **425** includes conversion rules for converting the universal glycan format **400** into the second platform-specific glycan format **420**. In some embodiments, the third rule set **435** includes conversion rules for converting the third platform-specific glycan format **430** into the universal glycan format **400**. In some embodiments, the third rule set **425** includes conversion rules for converting the universal glycan format **400** into the third platform-specific glycan format **430**. In some embodiments, the N^{th} rule set **445** includes conversion rules for converting the N^{th} platform-specific glycan format **440** into the universal glycan format **400**. In some embodiments, the N^{th} rule set **445** includes conversion rules for converting the universal glycan format **400** into the N^{th} platform-specific glycan format **440**.

[0075] A glycan data management system of the present disclosure can include any suitable number of format conversion rule sets in the plurality of format conversion rule sets, where each distinct rule set provides for conversion of a distinct platform-specific glycan format to and from the universal glycan format. In some embodiments, the number of format conversion rule sets in the plurality of format conversion rule sets is 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50, 60, 70, 80, 90, 100 or more, or a number in a range defined by any two of the preceding values. In some embodiments, the number of rule sets R in the plurality of format conversion rule sets is equal to the number of platform-specific glycan formats, e.g., the number of distinct platform-specific glycan formats.

[0076] In some embodiments, the instructions do not include a rule set for directly converting the input glycan data between different platform-specific glycan formats. In some embodiments, the instructions do not include a rule set for directly converting the input glycan data between the first platform-specific glycan format and the second platform-specific glycan format. In some embodiments, the instructions do not

include a rule set for directly converting the input glycan data between the first platform-specific glycan format and the third platform-specific glycan format. In some embodiments, the instructions do not include a rule set for directly converting the input glycan data between the second platform-specific glycan format and the third platform-specific glycan format. In some embodiments, the instructions do not include a rule set for directly converting the input glycan data between at least two of the N platform-specific glycan formats. In some embodiments, the instructions do not include a rule set for directly converting the input glycan data between any two of the N platform-specific glycan formats.

[0077] In some embodiments, the instructions include a rule set for directly converting the input glycan data between different platform-specific glycan formats. In some embodiments, the instructions include a rule set for directly converting the input glycan data between the first platform-specific glycan format and the second platform-specific glycan format. In some embodiments, the instructions include a rule set for directly converting the input glycan data between the first platform-specific glycan format and the third platform-specific glycan format. In some embodiments, the instructions include a rule set for directly converting the input glycan data between the second platform-specific glycan format and the third platform-specific glycan format. In some embodiments, the instructions include a rule set for directly converting the input glycan data between at least two of the first to N^{th} platform-specific glycan formats. In some embodiments, the instructions include a rule set for directly converting the input glycan data between any two of the N platform-specific glycan formats.

[0078] In some embodiments, a first glycan format includes glycan structural information, and the rule set is configured to convert the first glycan format to and from a second glycan format that does not include glycan structural information (e.g., a format that includes compositional information only). In some embodiments, a glycan representation in a first format, e.g., a first platform-compatible format or a universal glycan format, having glycan structural information can be converted into a second format, e.g., a second platform-compatible format or a universal glycan format, that includes compositional information only (and no structural information) by enumerating the amount of each of the monosaccharides in the glycan. In some embodiments, enumerating the amount of each of the monosaccharides in the glycan data includes counting the number of each type of monosaccharide in the glycan, and representing the total number for each type of monosaccharide in the glycan. In some embodiments, converting a glycan

representation in a first format, e.g., a first platform-compatible format or a universal glycan format, that does not have glycan structural information, to a second format, e.g., a second platform-compatible format, that includes structural information, includes receiving information about a source of the glycan data; and generating a list of potential glycan structures based on the compositional information and the source of the glycan data. Without being bound by theory, the possible glycan structures that may be present in a sample can depend on the source of the sample. For example, for a given composition, the list of potential glycan structures in human serum is different from the list of potential glycan structures in human stool. As such, in some embodiments, generating the list of potential glycan structures includes determining the source of the sample and determining the possible or likely glycan structures for the sample. The potential structures may then be listed as conversions of the first format to the second format in this example.

[0079] Also provided herein are methods, e.g., computer-implemented methods, and conversion rule sets for converting between glycan formats that include structural information and those that do not include structural information, e.g., those that include compositional information only. With reference to Fig. 6, a non-limiting method **600** of converting a glycan format, e.g., between formats that include and do not include glycan structural information, is provided. The method can include receiving **610** glycan data comprising a glycan representation in a first format that includes compositional information. The method can further include converting **620** the glycan representation to a second format that includes compositional information, where either: (1) the first format further comprises glycan structural information, and the second format comprises compositional information only, and wherein converting the glycan representation comprises enumerating an amount of each of the monosaccharides in the glycan data; or (2) the first format comprises compositional information only, and the second format further comprises glycan structural information, and wherein converting the representation comprises: receiving information about a source of the glycan data; and generating a list of potential glycan structures based on the compositional information and the source of the glycan data. The method can be implemented in any suitable system, such as a glycan data management system of the present disclosure.

[0080] The source of the glycan data can by any suitable biological or environmental source. In some embodiments, the glycan data is from a glycan database that lists the potential glycan structures that are relevant for the type of sample from which glycomolecule mass spectrometry data is obtained and for which a glycan search is being

performed. In some embodiments, the source includes, without limitation, urine, stool, blood, or serum, e.g., human urine, human stool, human blood, or human serum. In some embodiments, the source includes yeast, a plant, a non-human mammal, or a human.

Glycan databases

[0081] The glycan database for use in the systems and methods of the present disclosure can be any suitable database of glycan data. In general, a glycan database serves as an input to a glycomolecule search engine, e.g., a glycopeptide search engine. In some embodiments, the glycan database includes representations of glycan data for one or more glycans. The glycan database can include representations of glycan data in any suitable glycan format (e.g., platform-specific glycan format, universal glycan format), as disclosed herein. In some embodiments, the glycan database includes representations of glycan data for 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 10^4 ,10^5, 10^6 or more, or for a number in a range defined by any two of the preceding values, glycans. In some embodiments, the glycan database is specific for the type of biological sample or source of the glycans. In some embodiments, the source or biological sample is blood, serum, urine, or stool. In some embodiments, the source is a human, a non-human mammal, a plant or yeast, or the biological sample is from a human, a non-human mammal, a plant or yeast.

[0082] The glycan database for use in the systems and methods of the present disclosure can be provided to the system or in the method by any suitable option. In some embodiments, the glycan database is from a third-party source. In some embodiments, the glycan database is obtained from a remote server, such as a third-party server, e.g., via a network interface of the system. In some embodiments, the glycan database is stored locally, e.g., in a data storage device of the system, or a local server. In some embodiments, the glycan database is provided by the user, e.g., by typing in or copy-pasting glycan representations into an interface.

[0083] In some embodiments, the glycan database is compatible with a glycomolecule search engine, e.g., can be an input to a glycomolecule search engine, as disclosed herein. In some embodiments, the glycan database is compatible with a glycopeptide search engine, such as, BYONIC, PGLYCO (e.g., PGLYCO3), or METAMORPHEUS.

Additional embodiments

[0084] In some embodiments, the system includes a Glycopeptide Data Converter, as shown in Fig. 1, which can take the results from each database search and convert those results into a universal output format. It should be realized that the output from each database may include a confidence score for each result which can also be converted into the universal glycan format so that each output, and its confidence value, can be compared. Thus, embodiments provide a glycan database converter for conversion of glycan database formats (the inputs), and a novel glycopeptide result converter for conversion of glycopeptide results from different search engines (the outputs).

[0085] As described below, these converters were applied on a cohort of human serum samples and reported confidence score dynamics for different N-linked glycopeptide search engines, including PGLYCO3, BYONIC, and METAMORPHEUS.

[0061] With reference to Fig. 11, non-limiting embodiments of glycomolecule data analysis systems and methods for comparing glycomolecule search outputs are provided. A glycomolecule data analysis system can include a processor, e.g., a hardware processor such as a central processing unit (CPU); and a non-transitory memory configured to cause the processor to (e.g., having instructions thereon that, when executed, cause the processor to): access a first search output associated with a first glycomolecule search engine **1111** and a second search output associated with a second glycomolecule search engine **1112**, wherein the first search output comprises first glycomolecular data for each of one or more identified first glycomolecules in a first format **1121**, and wherein the second search output comprises second glycomolecular data for each of one or more identified second glycomolecules in a second format **1122**. The non-transitory memory can further be configured to cause the processor to (e.g., have instructions thereon that, when executed, cause the processor to): convert one or more (e.g., both) of the first and second glycomolecular data to a common format **1131**, **1132**, wherein the common format comprises the first format, the second format, or a universal glycomolecular data format. The non-transitory memory can further be configured to cause the processor to (e.g., have instructions thereon that, when executed, cause the processor to): perform a comparison between the first and second glycomolecular data in the common format. In some embodiments, the non-transitory memory is further configured to cause the processor to: access a third search output associated with a third glycomolecule search engine, wherein the third search output comprises third glycomolecular data for each of one or more

identified third glycomolecules in a third format; convert one or more of the first, second, and third glycomolecular data to a common format, wherein the common format comprises the first format, the second format, the third format, or the universal glycomolecular data format; and perform a comparison between any two or more (e.g., between all three) of the first, second, and third glycomolecular data, the comparison performed in the common format. In some embodiments, the non-transitory memory is further configured to cause the processor to: access an N^{th} search output associated with an N^{th} glycomolecule search engine **1113**, wherein the N^{th} search output comprises N^{th} glycomolecular data for each of one or more identified N^{th} glycomolecules in an N^{th} format **1123**; convert one or more (e.g., two or more, or all N) of the first to N^{th} glycomolecular data to a common format (e.g., **1131**, **1132**, **1133**), wherein the common format comprises any one of the first to N^{th} format, or the universal glycomolecular data format; and perform a comparison between any two or more of the first to N^{th} glycomolecular data, the comparison performed in the common format, where N is an integer that is two or greater.

[0062] The glycomolecule search engines can generate the search outputs based on a suitable input data set **1150**, which can include glycomolecule sample data, e.g., mass spectrometry data. In some embodiments, the search outputs accessed or retrieved in the systems and methods of the present disclosure are those obtained from different glycomolecule search engines that are run on the same glycomolecule sample data, e.g., mass spectrometry data in the input data set. In some embodiments, the input data set provided to each glycomolecule search engine, from which a search output is obtained, includes the same glycomolecule sample data, e.g., obtained from the same experiment or mass spectrometry run.

[0063] The glycomolecular data in the search output can provide information about a glycomolecule that the search engine identifies in the input data set. In general, the glycomolecular data includes information about a glycan identified in the sample, e.g., glycan data, based on the glycomolecule sample data, e.g., mass spectrometry data, in the input data set. In some embodiments, the glycan data includes glycan compositional information. In some embodiments, the glycan data includes glycan structural information. In some embodiments, the glycan data does not include glycan structural information. In some embodiments, the glycan data includes glycan compositional information only. In some embodiments, the glycan data includes the amount of hexose, N-acetylhexosamine, fucose, and/or sialic acid (e.g., N-acetylneuraminic acid, N-glycolylneuraminic acid) associated with a glycomolecule.

[0064] In some embodiments, the glycomolecular data includes one or more parameters associated with each identified glycomolecule in the data. In some embodiments, the glycomolecular data includes, without limitation, one or more of: glycan data, an identifier (in the field of mass spectrometry for example, the identifier is sometimes referred to as a "scan number"), a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score. The identifier can be used to identify an analyte for which sample data, e.g., mass spectrometry data, was obtained, and can serve as a label to identify the same analyte in search outputs obtained from different glycomolecule search engines, e.g., glycopeptide search engines, such as, without limitation, PGLYCO, BYONIC, METAMORPHEUS, STRUCGP, etc., that are run on the same mass spectrometry data. In some embodiments, the glycan data provides compositional and/or structural information for glycans determined as being likely present in a sample (e.g., glycans of glycopeptides, glycolipids, glycoRNA, glycoDNA) based on the sample data, e.g., mass spectrometry data. In embodiments where a glycopeptide search is performed, the search output may include glycan data and the peptide sequence of the polypeptide to which a glycan is conjugated, based on the sample data. In some embodiments the retention time, the m/z value, and/or the charge number are provided for the analyte (e.g., a glycopeptide) that is identified based on the sample data. In some embodiments, the confidence score provides an indication of how well the data for a particular analyte matches the glycomolecule identified by the search engine. In some embodiments, the confidence score in the search output is specific to the search engine. In some embodiments, the confidence score provides an indication of how well the data for a particular analyte matches the entire glycomolecule (e.g., the glycan and the non-glycan component of the glycomolecule) identified by the search engine. In some embodiments, the confidence score provides an indication of how well the data for a particular analyte matches the non-glycan component of the glycomolecule identified by the search engine. The confidence score may be determined by one or more algorithms of a respective search engine. In some embodiments, where the glycomolecule is a glycopeptide, the confidence score provides an indication of how well the data for a particular analyte matches the peptide component of the glycopeptide identified by the search engine. Due to differences in search engine algorithms and formats, different search engines may return different confidence scores for the same analyte such that search output confidence scores cannot be easily matched or compared among search outputs from different search engines. This is discussed in further detail below.

[0065] In some embodiments, the glycomolecular data (e.g., the glycan data and/or the parameters associated with each identified glycomolecule, as disclosed herein) in the search output are in a search engine-specific format (e.g., **1121, 1122, 1123**) that is associated with the search engine corresponding to the search output. In these embodiments, when a search output from a particular search engine is accessed or retrieved by a system or in a method of the present disclosure, the search output may be in a search engine-specific format that is associated with the particular search engine. For example, the glycomolecular data (e.g., any one or all the parameters associated with each identified glycomolecule) in the search output from the PGLYCO3 search engine is in a PGLYCO-specific format. As another example, the glycomolecular data (e.g., any one or all the parameters associated with each identified glycomolecule) in the search output from the BYONIC search engine is in a BYONIC -specific format. In some embodiments, the glycomolecular data (e.g., any one or all the parameters associated with each identified glycomolecule) in the search output from METAMORPHEUS search engine is in a METAMORPHEUS-specific format. In some embodiments, the glycomolecular data (e.g., any one or all the parameters associated with each identified glycomolecule) in the search output from the STRUCGP search engine is in a STRUCGP-specific format. In some embodiments, the identifier in the search output is in a search engine-specific format (e.g., a PGLYCO-specific format, BYONIC-specific format, METAMORPHEUS-specific format, or STRUCGP-specific format). In some embodiments, the peptide sequence in the search output is in a search engine-specific format (e.g., a PGLYCO-specific format, BYONIC-specific format, METAMORPHEUS-specific format, or STRUCGP-specific format). In some embodiments, the protein ID in the search output is in a search engine-specific format (e.g., a PGLYCO-specific format, BYONIC-specific format, METAMORPHEUS-specific format, or STRUCGP-specific format). In some embodiments, the retention time in the search output is in a search engine-specific format (e.g., a PGLYCO-specific format, BYONIC-specific format, METAMORPHEUS-specific format, or STRUCGP-specific format). In some embodiments, the m/z value in the search output is in a search engine-specific format (e.g., a PGLYCO-specific format, BYONIC-specific format, METAMORPHEUS-specific format, or STRUCGP-specific format). In some embodiments, the charge number in the search output is in a search engine-specific format (e.g., a PGLYCO-specific format, BYONIC-specific format, METAMORPHEUS-specific format, or STRUCGP-specific format). In some embodiments, the confidence score in the search output is in a search engine-specific format (e.g., a PGLYCO-specific

format, BYONIC-specific format, METAMORPHEUS-specific format, or STRUCGP-specific format).

[0066] In some embodiments, the disclosed system may be configured to automatically detect a format of the search output. For example, the system may compare the text or arrangement of the glycomolecular data to known formats to determine that the search output matches a particular format. In some embodiments, the user may submit a user input selecting a particular format to indicate that an associated search output is in the particular format.

[0067] In some embodiments, the glycomolecular data (e.g., glycan compositional and/or structural information) in the search output is in a search engine-specific format, such as when the search output is accessed or retrieved by a system or in a method of the present disclosure. In some embodiments, the glycan data in the search output from PGLYCO (e.g., PGLYCO3) is in a PGLYCO-specific format. In some embodiments, the glycan data in the search output from BYONIC is in a BYONIC-specific format. In some embodiments, the glycan data in the search output from METAMORPHEUS is in a METAMORPHEUS-specific format. In some embodiments, the glycan data in the search output from STRUCGP is in a STRUCGP-specific format.

[0068] In some embodiments, a number of searches may be conducted on multiple search engines, resulting in multiple search outputs having glycomolecular data (e.g., the first, second, and/or third glycomolecular data) of different search engine-specific formats. The glycomolecular data from the multiple search outputs may be converted to a common format (e.g., **1131**, **1132**, **1133**). That is, the data may all be converted to one of the search engine-specific formats such that they are all in the same format. In some embodiments, where the common format is a search engine-specific format, the glycomolecular data is converted from a search engine-specific format that is not the common format to the search engine-specific format that is the common format. In some embodiments, the common format is a search output format employed by one of the glycomolecule search engines, e.g., glycopeptide search engines, such as, without limitation, PGLYCO, BYONIC, METAMORPHEUS, STRUCGP, etc. In some embodiments, the common format is a PGLYCO-specific format, BYONIC-specific format, METAMORPHEUS-specific format, or a STRUCGP-specific format. In some embodiments, the common format is not the PGLYCO-specific format. In some embodiments, the common format is not the BYONIC-specific format. In some embodiments, the common format is not the

METAMORPHEUS-specific format. In some embodiments, the common format is not the STRUCGP-specific format.

[0069] In some embodiments, the glycomolecular data, e.g., the first, second, and/or third glycomolecular data, is converted to a common format (e.g., **1131**, **1132**, **1133**) that is a universal glycomolecular data format. The universal glycomolecular data format is described in greater detail below. In some embodiments, the common format is not a search engine-specific format (e.g., not the PGLYCO-specific format, BYONIC-specific format, METAMORPHEUS-specific format, or the STRUCGP-specific format).

[0070] The glycomolecular data, e.g., the first, second, and/or third glycomolecular data, can be converted to the common format using any suitable option. In some embodiments, the glycomolecular data, e.g., the first, second, and/or third glycomolecular data, is converted to the common format by: extracting one or more parameters, as disclosed herein, associated with each of the one or more identified glycomolecules (e.g., identified first, second and/or third glycomolecules) in the glycomolecular data (e.g., first, second and/or third glycomolecular data); and converting the one or more parameters to the common format. The parameters can include one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score. In some embodiments, an output that includes the glycomolecular data (e.g., first, second and/or third glycomolecular data) in the common format can be generated as an output (e.g., a new data file, a printed output, displayed on a screen, etc.). In some embodiments, the common format is a universal glycomolecular data format, as disclosed herein. In some embodiments, the glycomolecular data, e.g., the first, second, and/or third glycomolecular data, is converted to the universal glycomolecular data format, as disclosed herein.

[0071] In some embodiments, the glycomolecular data, e.g., the first, second, and/or third glycomolecular data, is converted to a common format by removing one or more parameters in a glycomolecule search engine-specific format. In some embodiments, the glycomolecular data, e.g., the first, second, and/or third glycomolecular data, is converted to a common format by removing one or more parameters that is unique to search output from a particular glycomolecule search engine. In some embodiments, the first glycomolecular data is converted to a common format by removing one or more parameters that is present in search output from the first glycomolecule search engine but is absent from search output from a second glycomolecule search engine. In some embodiments, the first glycomolecular data is converted to a common format by removing one or more

parameters that is present in search output from the first glycomolecule search engine but cannot be derived from the search output from a second glycomolecule search engine.

[0072] In some embodiments, converting the glycomolecular data, e.g., the first, second, and/or third glycomolecular data, to a common format involves deriving a consensus confidence score from one or more search engine-specific confidence scores associated with the identified glycomolecule, as described in further detail below.

[0073] Also provided is a glycomolecule data analysis system for comparing glycomolecule search outputs from two or more glycomolecule search engines, where the system includes a processor, e.g., a hardware processor such as a central processing unit (CPU); and a non-transitory memory configured to cause the processor to (e.g., having instructions thereon that, when executed, cause the processor to): access two or more search outputs from two or more glycomolecule search engines, each search output (1) being associated with one of the two or more glycomolecule search engines and (2) comprising glycomolecular data in a search engine-specific format for each of one or more glycomolecules identified by the glycomolecule search engine. The non-transitory memory can further be configured to cause the processor to (e.g., have instructions thereon that, when executed, cause the processor to): extract one or more parameters (e.g., as discussed herein) associated with each of the one or more identified glycomolecules in the glycomolecular data of each of the two or more search outputs, where the one or more parameters includes an identifier. The non-transitory memory can further be configured to cause the processor to (e.g., have instructions thereon that, when executed, cause the processor to): perform a comparison between the glycomolecular data from the two or more search outputs based on the one or more extracted parameters. As discussed above, the identifier provides a label to identify the same analyte in search outputs from different search engines when they are run on the same sample data, e.g., the same mass spectrometry data. In some embodiments, the comparison is performed by identifying glycomolecules associated with the same identifier in the glycomolecular data from the two or more search outputs. In some embodiments, the extracted parameter further includes one or more additional features of the glycomolecular data. Suitable extracted parameters include, without limitation, one or more of: glycan data, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score. The search outputs accessed can be any suitable search outputs, as disclosed herein, e.g., from two or more glycomolecule search engines that are run on the same glycomolecule sample data, e.g., mass spectrometry data.

[0074] In some embodiments, the non-transitory memory is further configured to cause the processor to convert glycomolecular data of one or more of the two or more search outputs to a common format, such as a particular search engine-specific format or a universal glycomolecular data format, as disclosed herein. In some embodiments, the common format is the universal glycomolecular data format. In some embodiments, the converted search outputs are combined to generate a combined database that includes glycomolecular data from two or more of the search outputs in the common format, e.g., in the universal glycomolecular data format.

[0075] As disclosed herein, the use of a common format to which search outputs from different glycomolecule search engines are converted can facilitate comparison of the search outputs across different search engine platforms (e.g., PGLYCO, BYONIC, METAMORPHEUS, STRUCGP), including, but not limited to, when the search engines are run on the same glycomolecule mass spectrometry data. In some embodiments, where the search outputs are obtained from different search engines run on the same glycomolecule mass spectrometry data, e.g., from the same experiment or mass spectrometry run, different search outputs are compared on an analyte basis. In some embodiments, glycomolecular data for the same identifier is compared across different search outputs from different search engines. By doing so, glycomolecular data determined by the different search engines for the same analyte may be compared. As explained above, this comparison may allow for validation or corroboration of the identity of analytes in a sample from mass spectrometry data.

[0076] In some embodiments, the comparison includes identifying one or more overlaps between any two or more of the glycomolecular data. Without being bound by theory, overlap between search results from different glycomolecule or glycopeptide search engines can in some cases indicate which of the identified glycomolecules or glycopeptides are more reliable identifications. In general terms, an overlap is found between different glycomolecular data when there is agreement in identification of one or more parameters of the glycomolecular data for the same analyte, e.g., for identified glycomolecules having the same identifier. In some embodiments, the overlap includes a set of glycomolecules that are identified in any two or more of the glycomolecular data. In some embodiments, the overlap includes a set of glycomolecules having the same glycan composition in different search outputs, e.g., different identified glycomolecules associated with the same identifier. In some embodiments, the overlap includes a set of glycomolecules having the same glycan composition and peptide sequence in different search outputs, e.g., different

identified glycomolecules associated with the same identifier. In some embodiments, the overlap includes one or more of peptide sequences, protein IDs, retention times, m/z values, or charge numbers.

[0077] In some embodiments, the comparison includes identifying a difference between any two or more of the glycomolecular data. Without being bound by theory, differences between search results from different glycomolecule or glycopeptide search engines can in some cases indicate which of the identified glycomolecules or glycopeptides are less reliable identifications. In general terms, a difference is identified between different glycomolecular data when there is disagreement as to the identity or a property of an analyte. For example, a first search engine may identify (e.g., to a threshold confidence level) a first analyte as a first glycopeptide while a second search engine may identify the same analyte as a second glycopeptide that is different from the first glycopeptide (e.g., to a threshold confidence level). In some embodiments, the difference includes one or more sets of glycomolecules that are identified in one of the search results but not in another. In some embodiments, the difference includes one or more sets of glycomolecules having different glycan compositions in different search outputs, e.g., different identified glycomolecules associated with the same identifier. In some embodiments, the difference includes one or more sets of glycomolecules having different glycan compositions and peptide sequences in different search outputs, e.g., different identified glycomolecules associated with the same identifier. In some embodiments, the overlap includes one or more of peptide sequences, protein IDs, retention times, m/z values, or charge numbers. In some embodiments, the difference includes one or more of peptide sequences, protein IDs, retention times, m/z values, or charge numbers.

[0078] Also provided herein is a method of comparing glycomolecule search outputs from two or more glycomolecule search engines. With reference to Fig. 12, the method **1200** can include receiving **1210** a first search output associated with a first glycomolecule search engine and a second search output associated with a second glycomolecule search engine, wherein the first search output comprises first glycomolecular data for each of one or more identified first glycomolecules in a first format, and wherein the second search output comprises second glycomolecular data for each of one or more identified second glycomolecules in a second format. The method can further include converting **1220** one or more of the first and second glycomolecular data to a common format, wherein the common format comprises the first format, the second format, or a universal glycomolecular data format. The method can also include comparing

1230 the first and second glycomolecular data in the common format. The method can include receiving search outputs from any suitable number of glycomolecule search engines. In some embodiments, the method includes receiving a third search output associated with a third glycomolecule search engine, where the third search output comprises third glycomolecular data for each of one or more identified third glycomolecules in a third format; converting one or more of the first, second, and third glycomolecular data to a common format, wherein the common format comprises the first format, the second format, the third format, or the universal glycomolecular data format; and comparing any two or more of the first, second, and third glycomolecular data, each in the common format. In some embodiments, the method includes receiving a N^{th} search output associated with a N^{th} glycomolecule search engine, where the N^{th} search output comprises N^{th} glycomolecular data for each of one or more identified N^{th} glycomolecules in a N^{th} format; converting one or more of the first to N^{th} glycomolecular data to a common format, wherein the common format comprises any one of the first to N^{th} format, or the universal glycomolecular data format; and comparing any two or more (e.g., two or more, all N) of the first to N^{th} glycomolecular data, each in the common format, where N is an integer 2 or greater.

[0079] In some embodiments, the common format is the universal glycomolecular data format as disclosed herein. In some embodiments, the common format is a search engine-specific format.

[0080] In some embodiments, the glycomolecular data includes one or more parameters associated with each of the glycomolecules identified in the glycomolecular data, where the parameters are in a search engine-specific format and includes one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score. Any one or more of the parameters can be converted to the common format.

[0081] In some embodiments, the method includes providing a data set for running on each of the two or more glycomolecule search engines, wherein the data sets include the same glycomolecule sample data, e.g., mass spectrometry data. In some embodiments, the glycomolecule sample data, e.g., mass spectrometry data, in the data sets provided to the different glycomolecule search engines are obtained from the same experiment or glycomolecule mass spectrometry run. In some embodiments, the method includes generating the glycomolecule sample data from a biological sample, e.g., by running a glycomolecule mass spectrometry experiment on the sample; and providing the data set

containing the generated glycomolecule sample data. The data set can include any suitable component of an input to a glycomolecule search engine. In some embodiments, the data set further includes, without limitation, one or more of a glycan database and a protein/peptide database (or any other suitable database, e.g., a lipid database, a RNA database, a DNA database).

[0082] In some embodiments, the two or more glycomolecule search engines are two or more glycopeptide search engines, and the method further includes providing a dataset for running on each of the two or more glycopeptide search engines, where the datasets include the same glycopeptide sample data, e.g., glycopeptide sample data from the same experiment or mass spectrometry run.

[0083] With reference to Fig. 13, also provided herein is a method **1300** for determining a glycomolecule profile of a biological sample. The method can include providing **1310** to each of two or more glycomolecule search engines a data set comprising glycomolecule sample data, e.g., mass spectrometry data, obtained from a biological sample, wherein the same glycomolecule sample data, e.g., mass spectrometry data, are provided to each of the two or more glycomolecule search engines. The method can further include receiving **1320** an output from each of the two or more glycomolecule search engines run on the provided data set, wherein each output from the two or more glycomolecule search engines comprises glycomolecular data for each of one or more identified glycomolecules in a search engine-specific format. The method can include converting **1330** glycomolecular data of one or more of the outputs from the two or more glycomolecule search engines to a common format, wherein the common format comprises one of the search engine-specific formats or a universal glycomolecular data format. The method can also include determining **1340** a glycomolecule profile of the sample based on the glycomolecular data of the outputs from the two or more glycomolecule search engines in the common format. In some embodiments, the common format is the universal glycomolecular data format. In some embodiments, the common format is a search engine-specific format, as provided herein.

[0084] The glycomolecule profile can be determined based on the glycomolecular data of the glycomolecule search engines in the common format using any suitable option. In some embodiments, determining the glycomolecule profile comprises generating a consensus list of glycomolecules identified in the glycomolecular data associated with at least two of the glycomolecule search engines.

[0085] In some embodiments, each data set includes a glycan database containing glycan data in a glycan format that is compatible with a corresponding glycomolecule search engine of the two or more glycomolecule search engines, e.g., a corresponding glycopeptide search engine of the two or more glycopeptide search engines. The glycan data can be in any suitable glycan format that is compatible with, or is used by, a glycomolecule search engine. In some embodiments, the glycan format is compatible with a glycopeptide search engine such as, but not limited to, BYONIC, PGLYCO (e.g., PGLYCO3), or METAMORPHEUS search engines.

[0086] Converting glycomolecular data in one format, e.g., a search engine-specific format, to a common format can be done using any suitable option. In some embodiments, converting glycomolecular data to a common format includes extracting the one or more parameters, as disclosed herein, associated with each of the one or more identified glycomolecules in the glycomolecular data; and converting the one or more parameters to the common format. In some embodiments, the common format is a universal glycomolecular data format, as disclosed herein. In some embodiments, the converted glycomolecular data of the search outputs are combined to generate a combined database that includes glycomolecular data from two or more of the search outputs in the common format, e.g., in the universal glycomolecular data format.

[0087] In any method of the present disclosure, each of the glycomolecular data can include any suitable number of identified glycomolecules. In some embodiments, the number of identified glycomolecules is 100 or more, 200 or more, 500 or more, 1,000 or more, 2,000 or more, 5,000 or more, 10,000 or more, 20,000 or more, 50,000 or more, 100,000 or more, 500,000 or more, 10^6 or more, or a number in a range defined by any two of the preceding values.

[0088] In any method of the present disclosure, the method can involve search outputs from any suitable number of glycomolecule search engines. In some embodiments, the number of glycomolecule search engines is 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50, 60, 70, 80, 90, 100 or more, or a number in a range defined by any two of the preceding values.

[0089] In any system or method of the present disclosure, in some embodiments, the glycomolecule is a glycopeptide, glycoprotein, glycolipid, glycoRNA, or glycoDNA. In some embodiments, the glycomolecule is a glycopeptide. In any system or method of the present disclosure, in some embodiments, the glycomolecule search engine is a glycopeptide search engine, glycoprotein search engine, glycolipid search engine,

glycoRNA search engine, or glycoDNA search engine. In some embodiments, the glycomolecule search engine is a glycopeptide search engine. The glycomolecule search engine can be any suitable software platform configured to receive input data, e.g., glycomolecule sample data or mass spectrometry data, perform a search of the input data against a suitable database, and generate an output of the search results, e.g., with glycomolecular data. In some embodiments, the glycomolecule search engine is a publicly available search engine. In some embodiments, the glycomolecule search engine is a commercially available search engine. In some embodiments, the glycomolecule search engine is a third-party search engine.

[0090] The sample or biological sample in any system or method of the present disclosure can be any suitable biological source of glycomolecules (e.g., glycopeptide, glycoprotein, glycolipid, glycoRNA, glycoDNA). In some embodiments, the biological sample is a urine, stool, blood, or serum sample. In some embodiments, the sample is from human urine, human stool, human blood, or human serum. In some embodiments, the sample is from yeast, a plant, a non-human mammal, or a human. In some embodiments, a method of the present disclosure includes generating the glycomolecule sample data or mass spectrometry data from the biological sample.

[0091] In any system or method of the present disclosure, in some embodiments, the results of the comparison can be further processed in any suitable manner for the intended use. In some embodiments, the results of the comparison are provided as an output, e.g., in the common format, for the user. In some embodiments, the results of the comparison are provided on a screen, as an electronic file, as a printout, an audio output, etc. In some embodiments, a report is generated with the results of the comparison. In some embodiments, the results of the comparison are stored in a database. In some embodiments, the results of the comparison from different experiments (e.g., having different glycomolecule sample data or mass spectrometry data) are combined to generate a combined output. In some embodiments, the results of the comparison from different experiments (e.g., having different glycomolecule sample data or mass spectrometry data) are stored in the same database.

Universal glycomolecular data format

[0092] In any system or method of the present disclosure, the universal glycomolecular data format can include any suitable representation of glycomolecular data that conveys information desired and/or parameters extracted from search outputs from two

or more glycomolecule search engines. As disclosed herein, the universal glycomolecular data format can facilitate comparison of the search results from two or more glycomolecule search engines. In some embodiments, the universal glycomolecular data format includes, for each identified glycomolecule in the glycomolecular data, without limitation, one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score.

[0093] In some embodiments, the universal glycomolecular data format includes glycan data that includes representations of glycans that include compositional and/or structural information, of the glycans identified as being associated with a particular glycomolecule. In some embodiments, the glycan data includes glycan compositional information. In some embodiments, the glycan data includes the amount of hexose, N-acetylhexosamine, fucose, and/or sialic acid (e.g., N-acetylneuraminic acid, N-glycolylneuraminic acid) associated with a glycomolecule. In some embodiments, the glycan data includes glycan structural information. In some embodiments, the glycan data does not include glycan structural information. In some embodiments, the glycan data includes glycan compositional information only. In some embodiments, the universal glycomolecular data format includes glycan data in more than one representation. In some embodiments, the universal glycomolecular data format includes glycan data in one or more representations with compositional information only, and in one or more representations with glycan structural information. The glycan data can be in any suitable format. In some embodiments, the glycan data in the universal glycomolecular data format is in a search engine-specific format, e.g., BYONIC, PGLYCO (e.g., PGLYCO3), STRUCGP, and/or METAMORPHEUS format. In some embodiments, the glycan data is in a universal glycan format. In some embodiments, the universal glycomolecular data format includes glycan data in more than one formats. In some embodiments, the universal glycomolecular data format includes glycan data in the universal glycan format and any one or more of the search engine-specific formats.

[0094] In some embodiments, the universal glycomolecular data format includes identifiers associated with each analyte detected and measured in an experiment. For example, an analyte (e.g., a glycopeptide) that passes through a mass spectrometer may be assigned a unique identifier, and the detected mass spectrum that results from this analyte passing through the mass spectrometer is associated with the assigned identifier. As disclosed herein, the identifier can be used to identify an analyte for which sample data (e.g., mass spectrometry data) was obtained, and can serve as a label to identify the same

analyte in search outputs obtained from different glycomolecule search engines, e.g., glycopeptide search engines, such as, without limitation, PGLYCO, BYONIC, METAMORPHEUS, STRUCGP, etc., that are run on the same sample data, e.g., mass spectrometry data. In some embodiments, the identifier is generally in a common format, or may otherwise be converted to a common format.

[0095] In some embodiments, the universal glycomolecular data format includes information about the molecules to which the glycans of each analyte are conjugated. For example, in the case where the analyte is a glycopeptide, the universal glycomolecular data format includes amino acid sequences of the peptides associated with the analyte (i.e., glycopeptides to which glycans are conjugated). In some embodiments, the peptide sequence is in a common format. In some embodiments, the peptide sequence in the universal glycomolecular data format represents the linear sequence of amino acids in the identified glycopeptide. In some embodiments, the peptide sequence in the universal glycomolecular data format includes one or more one or more amino acids flanking the identified glycopeptide in a glycoprotein from which the glycopeptide was taken. In other embodiments, the peptide sequence in the universal glycomolecular data format does not include any such amino acids flanking the identified glycopeptide in the glycoprotein. In some embodiments, the peptide sequence in the universal glycomolecular data format does not include a numerical character. In some embodiments, the peptide sequence in the universal glycomolecular data format includes any suitable abbreviation of amino acids. In some embodiments, the peptide sequence in the universal glycomolecular data format includes any suitable single-letter amino acid abbreviations. In some embodiments, the peptide sequence in the universal glycomolecular data format includes only the standard single-letter amino acid abbreviations. In some embodiments, the peptide sequence in the universal glycomolecular data format does not include non-standard single-letter amino acid abbreviations. In some embodiments, the peptide sequence in the universal glycomolecular data format indicates the site of glycosylation using a marker. For example, the marker includes an ASCII character, an image (e.g., an emoji), or some other suitable symbol positioned within the peptide sequence at the site of glycosylation (e.g., positioned before and/or after the amino acid to which a glycan is conjugated). As another example, the amino acid to which a glycan is conjugated may be formatted differently (e.g., it may be bolded, underlined, highlighted, colored differently). In some embodiments, the peptide sequence in the universal glycomolecular data format specifies the site of glycosylation

(e.g., by providing a numerical location within the peptide sequence or the protein sequence).

[0096] In some embodiments, the universal glycomolecular data format includes a protein ID. The protein ID can provide the identity of the protein which the glycopeptide identified in the glycopeptide sample data or mass spectrometry data is identified as being part of. The protein ID can be any suitable identifier used in a database of proteins, including, without limitation, UNIPROT, GENBANK, and ENSEMBL.

[0097] In some embodiments, the universal glycomolecular data format includes the retention time. The retention time can be in any suitable format. In some embodiments, the retention time is in minutes. In some embodiments, the retention time is in seconds.

[0098] In some embodiments, the universal glycomolecular data format includes the mass-to-charge (m/z) ratio for each analyte (e.g., corresponding to the ratio between the mass and charge of the precursor ion in a multiple reaction monitoring (MRM) experiment), which may be determined, for example, based on a mass spectrum associated with the analyte. The m/z value can be in any suitable format. For example, referencing Figs. 8 and 9, the numbers in the "PrecursorMZ" or "Observed m/z" column indicate the (m/z) ratio of each respective analyte. The m/z value can be rounded to any suitable number of decimal points. In some embodiments, the m/z value is rounded to 0, 1, 2, 3, 4, 5, 6, 7, 8, or more decimal points.

[0099] In some embodiments, the universal glycomolecular data format includes charge numbers for each analyte, indicating a charge associated with the respective analyte. The charge number can be in any suitable format. For example, referencing Figs. 8 and 9, the numbers in the "Charge" or "Z" column indicate the charge numbers of each respective analyte.

[0100] In some embodiments, the universal glycomolecular data format includes confidence scores for each analyte. The confidence score can be any suitable value that indicates how well the data matches the glycomolecule identified by a search engine. In other words, the confidence score may be an indication of a level of confidence the search engine's algorithm has that the correct glycomolecule was identified. In some embodiments, the confidence score of the universal glycomolecular data format is a confidence score selected from one of search outputs of a particular search engine (e.g., from one of the search engines that identified the glycomolecule based on the underlying data). In some embodiments, the format of the confidence score of the universal glycomolecular format may similarly be conformed to a particular one of the search

engines. For example, a first search engine-specific confidence score may be a number on a scale of 0 to 100, while a second search engine-specific confidence score may be a number on a scale of 0 to 1000. In this example, the confidence score of the universal glycomolecular format may be made to conform to the format of the first search engine (e.g., 0 to 100). In some embodiments, the format of the confidence score of the universal glycomolecular format may be a fixed format. For example, the confidence score of the universal glycomolecular format may always be on a scale from 0 to 100. In some embodiments, the confidence score for the universal glycomolecular data format is a consensus confidence score derived from the search engine-specific confidence scores of two or more search engines. For example, in the case of two search outputs from two search engines, for each analyte (corresponding to the same identifier in both search outputs), a consensus confidence score may be determined by applying a function to the two search engine-specific confidence scores of the two search outputs. For example, the consensus confidence score may be obtained by computing an average of the search engine-specific confidence scores for an analyte. As another example, the consensus confidence score may be obtained by computing a median of search engine-specific confidence scores an analyte. As another example, the consensus confidence score may be obtained by computing a weighted function (e.g., a weighted average) of search engine-specific confidence scores for an analyte. Computing a weighted function may include applying one or more different weighting factors to different search engines. In some embodiments, the confidence scores may first be normalized before applying a desired function (e.g., an average function, a median function, a weighted average function) to derive the consensus confidence score. For example, in the case of a first search engine-specific confidence score on a scale of 0 to 100 and a second search engine-specific confidence score on a scale of 0 to 1000, the scores may be normalized so that they are both on a scale of 0 to 100 or on a scale of 0 to 1000.

[0101] In some embodiments, the universal glycomolecular data format includes identifying information for the glycomolecule search engine. In general, the identifying information for the glycomolecule search engine can specify the search engine from which the search output was obtained. The identifying information for the glycomolecule search engine can be in any suitable format. In some embodiments, the identifying information for the glycomolecule search engine specifies whether the search engine is PGLYCO, BYONIC, METAMORPHEUS, or STRUCGP, or any other search engine from which the search output was obtained.

[0102] The universal glycomolecular data format can include one or more parameters for each of the identified glycomolecules in any suitable order. In some embodiments, the identifier is listed first. In some embodiments, the universal glycomolecular data format lists parameters in the following order: identifier, protein ID, peptide sequence, glycan data, m/z value, charge number, retention time, confidence score, search engine. In some embodiments, the universal glycomolecular data format lists parameters in the following order: identifier, protein ID, peptide sequence, glycan data, charge number, retention time, search engine.

[0103] Non-limiting examples of a universal glycomolecular data format are shown in Figs. 17-22.

[0037] FIG. 30 illustrates a schematic diagram showing a non-limiting example of a user interface presenting glycomolecule search results displayed in search result elements and a consensus set displayed in a consensus list. In some embodiments, a system for curating a consensus list of glycomolecules identified in a biological sample on a user interface is provided, the system including a processor and a memory (e.g., non-transitory memory) having instructions that, when executed by the processor, causes the processor to retrieve a plurality of glycomolecule search results sets for a biological sample from a plurality of glycomolecule search engines **3011a, b, ... *i*** (wherein $i \geq 3$, e.g., 3, 4, 5, 6, 7, 8, 9, 10, or more, hereafter). In some embodiments, retrieval of these search results may include accessing search results that were uploaded, imported, or otherwise inputted by the user into the system. In a non-limiting example, the system may be at least partially cloud-based, in which case the user may upload search results from multiple search engines (e.g., exported files from such search engines) via an online portal. The exported files may be in plain-text or formatted as a csv (comma-separated values), xml (extensible markup language), or any other suitable machine-readable format. In another non-limiting example, the system may be a local client device, in which case, the user may simply submit the search results as an input into the system. Such search engines include, but are not limited to, BYONIC, PGLYCO3, MSFRAGGER-GLYCO, METAMORPHEUS (also known as O-PAIR), STRUCGP, and other similar search engines can be used to match a mass spectrometry signature to known glycomolecule signatures. In some embodiments, each of the glycomolecule search results sets **3012a, b, . . . *i*** (wherein $i \geq 3$, e.g., 3, 4, 5, 6, 7, 8, 9, 10, or more, hereafter) includes glycomolecule search results **3014a, b, ... *i*** (wherein $i \geq 3$, e.g., 3, 4, 5, 6, 7, 8, 9, 10, or more, hereafter) that identify one or more predicted glycomolecules as being present. In some embodiments, the predictions are made based on

associated mass spectrometry data. In some of these embodiments, each predicted glycomolecule may be predicted based on an associated mass spectrum, which is itself associated with a unique identifier, typically called a mass spectrometry scan number. In some embodiments, the mass spectrum is a spectrum created by measuring signals from fragmented ions of a single type of glycomolecule that are sent through a mass spectrometer. For example, a multiple reaction monitoring setup may be used to ionize and filter for glycomolecules of a single type (e.g., filtered by retention times in a liquid chromatography column and the mass-to-charge ratio from a first mass spectrometry measurement), apply an energy to the ionized glycomolecules to fragment them into fragmented ions, and then record the spectrum of the fragmented ions using a second mass spectrometry measurement. In these embodiments, a unique identifier typically called an MS2 scan number is assigned to the spectrum of the fragmented ions, and each such MS2 scan number and corresponding spectrum ideally corresponds to a single type of glycomolecule. Referencing FIG. 30, the identifier **3013** may reference the MS2 scan number associated with a particular spectrum. In some embodiments, the instructions, when executed by the processor, further causes the processor to determine a consensus set which includes, for each unique identifier **3013**, a consensus search result identifying one of the predicted glycomolecules associated with the unique identifier **3013** in the glycomolecule search results sets. In one non-limiting example, as shown in FIG. 30, the consensus list element **3020** provides a glycan portion **3024a** of the predicted glycomolecule for Scan #1 (a glycomolecule, for example, may include a glycan and a peptide).

[0038] In some embodiments, the instructions, when executed by the processor, further causes the processor to provide a user interface **3002**, as shown in FIG. 30, configured to display one or more search results elements **3010a, b, ... *i*** (wherein $i \geq 3$, e.g., 3, 4, 5, 6, 7, 8, 9, 10, or more, hereafter) configured to display one or more representations of the one or more predicted glycomolecule search results **3014a, b, ... *i*** from each of the plurality of glycomolecule search results sets **3012a, b, ... *i***. In a non-limiting example, the glycomolecule search results set **3012a** from Search Engine 1 **3011a** is summarized and at least partially shown in the first search results element **3010a** as search results **3014a-1 ... 3014a-*n*** (wherein $n \geq 2$, e.g., 2, 3, 4, 5, 6, 7, 8, 9, 10, or more hereafter). In some embodiments, the representations of the one or more predicted glycomolecule search results **3014a, b, ... *i*** is textual, graphical, or is toggled between textual and graphical representations. In some embodiments, the one or more search results elements do not display the one or more representations of the one or more predicted glycomolecule search

results **3014a, b, … *i*** from each of the plurality of glycomolecule search results sets **3012a, b, … *i*** that are summarized in the search results elements **3010a, b, … *i*** respectively. In some embodiments, each of the predicted glycomolecule search results **3014a, b, … *i*** may be displayed in a plurality of respective search results elements **3010a, b, ... *i***. In some embodiments, information from all the glycomolecule search results or at least multiple glycomolecule search results is displayed in a single search result element (e.g., a single table element that displays all search results). In some embodiments, the respective locations of the plurality of search results elements **3010a, b, ... *i*** may be determined automatically on the user interface **3002** based on a default setting, or a user modification of the default setting that is read and executed by the processor. In some embodiments, the default setting may be based on a rank for the associated search engine, such rank being determined by the system disclosed herein. For example, the rank may be based on usage data (e.g., usage by the user or by a population of users) associated with each of the search engines. As another example, the rank may be based on a type of sample that is associated with the data that is being searched. That is, the system may determine that a first search engine is more optimal for a particular type of sample (e.g., as may be determined based on how commonly used it is for the particular type of sample) than a second search engine, and may thus rank the search engines accordingly. In such embodiments, the search results elements may be arranged in an order based on the rank associated with their respective search engines. In some embodiments, the user modification of the default setting takes precedence. In some embodiments, settings modifications such as this or any other suitable modification may be made via a user input element such as the settings element **3050**. In some embodiments, the user may be able reorder the arrangement of the search results elements on the user interface **3002** by submitting a suitable user input (e.g., by dragging and dropping or some other suitable means).

[0039] In some embodiments, the user interface **3002** is further configured to recognize a user input and allow scrolling through the one or more of the search results elements **3010a, b, ... *i***. In some embodiments, the user interface **3002** is configured such that scrolling activity in one of the search result elements **3010a** automatically causes the scrolling of the other search results elements **3010b ... *i*** such that sets of the glycomolecules associated with the same unique identifiers **3013** are displayed across all of the search results elements **3010a, b, ... *i*** as shown in FIG. 31A. As such, the user interface 3002 may enable the user to make direct comparisons among the different search results sets displayed in the search results elements **3010a, b, ... *i***. In some embodiments, the same set

includes the same representations. In some embodiments, the plurality of search results elements **3010a, b, ... *i*** may be arranged horizontally on the user interface as shown in FIG. 30, 31A-31B, and 32. In some embodiments, these plurality of search results elements **3010a, b, ... *i*** may be vertically arranged on the user interface as shown in FIG. 33. Any suitable arrangement is contemplated by this disclosure. In some embodiments, referencing FIGS. 32-33, the user interface **3002** is further configured to display a scroll bar **3042** that is activated and controlled by a user input, the scroll bar **3042** further permitting scrolling activity through the plurality of the search results elements **3010a, b, ... *i***, and thus, allow potentially large number of search results **3012** to be available for display on the user interface **3002**.

[0040] In some embodiments, referencing FIG. 30, the representations corresponding to the predicted glycomolecule search results **3014a, b, ... *i*** are displayed in each of the search result elements. In some embodiments, the unique identifiers **3013** associated with each search result is displayed to allow users to easily reference and compare search results across the different search results elements. Although FIG. 30 and the following figures only depict search results elements **3010a, b, ... *i*** that present scan numbers, compositional information of glycans, and confidence scores from search results, the disclosure contemplates that the search results elements **3010a, b, ... *i*** may also display other information from the search results, for example in additional columns. For example, in the case where the glycomolecule is a glycoconjugate that includes a glycan and another molecule bonded together (e.g., a glycopeptide that includes a glycan and a peptide, a glycoRNA that includes a glycan and an RNA, glycoDNA that includes a glycan and a DNA, a glycolipid that includes a glycan and a lipid), information such as sequence information of the other molecule may also be presented. As another example, information such as mass, mass-to-charge ration (m/z), and/or any other suitable information from the search results may also be presented in the search results elements **3010a, b, ... *i***. A consensus list element **3020** configured to display the consensus set **3022** may also be displayed. Consensus sets are described in further detail below.

[0041] In some embodiments, the search results **3012** include a confidence score **3015** for each of the glycomolecules in the conflict search results **3520**, and wherein the conflict resolution rule set **3540** includes identifying a consensus search result **3024** based on a comparison of the confidence scores **3015** of each search result among the conflicting search results **3520**. The confidence score can be any suitable value that indicates the degree of confidence a search engine may have that a predicted glycomolecule

search result corresponds to a particular mass spectrum. For example, referencing FIG. 30, Search Engine 1 reports a confidence score of 85 for its prediction that the mass spectrum associated with Scan #1 corresponds to its predicted glycomolecule search result (which includes the glycan N(1)H(1)(F)(1)A(0)). In some embodiments, the confidence score **3015** is a value that is reported in the search results of each and each value may thus be on its own respective search engine-specific scale. In some embodiments, the confidence score **3015** is a normalized and/or scaled confidence score that facilitates comparison of confidence scores across different search engines. For example, referencing FIG. 30, the confidence scores of two or more of the Search Engines 1 to *i* may have been scaled differently in the original search results that may have been exported from their respective search engines (e.g., Search Engine 1 may have been on a scale of 1 to 100, Search Engine 2 may have been on a scale of 1 to 500). But as shown in FIG. 30, the system may have scaled all the confidence scores to a common scale of 1 to 100 to allow for a more straightforward comparison. As another example, the confidence scores of two or more of the search engines may have different relative values. For example, a score of 61 reported by Search Engine 1 may have a relative value that is higher than a score of 65 reported by Search Engine 2, because the algorithm for Search Engine 1 may apply a different standard in scoring. In these cases, in some embodiments, the scores may be normalized by the disclosed system to account for this difference in relative values and adjusting the confidence scores. Normalization of the scores may be based on a predetermined weighting system that weights scores of different search engines differently, by manual input of a user, or may be dynamically determined based on for example historical performance of the different search engines.

[0042] FIG. 35 is a block diagram showing a non-limiting example of a method for populating a consensus list element **3020** according to some non-limiting embodiments of the disclosure. As shown in FIG. 35, in some embodiments, the consensus set is determined, at least in part, by identifying **3510** conflicting search results **3520** and non-conflicting search results **3530** among the predicted glycomolecule search results from different search engines, and populating the consensus list element **3020** with the non-conflicting search results **3530**. In some embodiments, the predicted glycomolecule search results are further evaluated by resolving the conflicting search results **3520** by the one or more programs that further include instructions, when executed by the processor, that applies a conflict resolution rule set **3540** to generate resolved conflicting search results **3550**, and populate the consensus list element **3020** with the resolved conflicting search

results **3550**. The conflict resolution rule set **3540** can include any suitable conflict resolution rules for resolving conflicting search results. In some embodiments, the conflict resolution rule set **3540** takes into account one or more factors associated with the search results, such as, without limitation, the number of search engines that are in agreement among a conflicting search result, the relative confidence scores (after normalization/scaling of the confidence scores) given by each search engine for the conflicting search result, etc. In some embodiments, the different factors are weighted differently to resolve a conflict. In some embodiments, the system is configured to apply a conflict resolution rule set **3540** that is a default or predetermined rule set (e.g., taking into account a default factor or set of factors, and/or to using a default weighting of the factors). In some embodiments, the user interface is configured to allow a user to determine which factor to use, and/or to select how different factors are to be weighted, for the conflict resolution rule set **3540**.

[0043] In some embodiments, the conflict resolution rule set **3540** comprises identifying, for a particular conflicting search result **3520** associated with a particular unique identifier **3013**, one or more plurality glycomolecules (e.g., a majority glycomolecule) from the predicted glycomolecule search results of the conflicting search results **3520** associated with the particular unique identifier **3013**. In some embodiments, a predicted glycomolecule is a plurality glycomolecule when the same glycomolecule is predicted by more search engines than the number of search engines identifying any one other glycomolecule. In some embodiments, a predicted glycomolecule is a majority glycomolecule when the same glycomolecule is predicted by more than half the search engines (e.g., by 2 out of 3 search engines, by 3 out of 4 search engines, by 3 or more out of five search engines, etc.). In some embodiments, resolving the conflict for a search result involves selecting one of the conflicting search results. For example, the system may select the plurality glycomolecule.

[0044] In some embodiments, again referencing FIG. 35, the search result among the conflicting search results **3520** having the highest confidence score is chosen as the resolved conflicting search results **3550** and therefore is populated **3560** in the consensus list element **3020** as shown in FIG. 35. In some embodiments, the conflicting search results **3520** are resolved based on relative confidence scores **3015** (e.g., calculating a weighted score for each potential glycomolecule). In a non-limiting example, as shown in FIG. 30 for the mass spectrum associated with Scan #4, two out of three search engines **3011a, b** may predict a glycomolecule (which includes the glycan N(4)H(4)F(3)A(0)) with

a combined confidence score of 70, but the remaining third search engine **3011*i***, may indicate another glycomolecule (which includes the glycan N(4)H(4)F(4)A(0)) as the predicted glycomolecule with a confidence score of 99. Because the confidence score from the third search engine **3011*i*** in this example is higher than the combined confidence score from the first and second search engines **3011a, b**, the predicted glycomolecule from the third search engine **3011*i*** may be selected and displayed on the consensus list element **3020** despite a majority of the search engines (2 out of 3 in this example) indicating a different glycomolecule. The confidence scores in this example may have been normalized and/or scaled by the system to a common scale of 1 to 100 as explained earlier. In another non-limiting example, confidence scores **3015** may be weighted based on known characteristics of the search algorithms employed by the search engines **3011** (e.g., some search engine algorithms are tuned to be better at classifying certain glycomolecules, better at analyzing certain sample types, etc.). In some embodiments, such characteristics may be based on historical performance metrics of each search engine **3011**. In some embodiments, users may be presented with characteristics/features that contributed to the confidence score **3015** for a particular scanned glycomolecule, which in some embodiments can advantageously assist the user in independent assessment of the confidence level of the associated prediction.

[0045] In some embodiments, determining the consensus set includes determining an initial consensus set, wherein the initial consensus set includes a single predicted glycomolecule search result for each unique identifier **3013**. In some embodiments, the consensus set **3022** includes an initial consensus set having subsets of the full set of glycomolecules identified in all the search results sets that are input into the disclosed system (e.g., the search results identified by all the search engines **3011**). In a non-limiting example, referencing FIG. 30, a user may input search results (e.g., displayed in search results elements 3011a, 3011b, and 3011*i*) from *i* search engines **3011** (e.g., Search Engine 1, Search Engine 2, . . . Search Engine *i*) to determine what glycomolecules may be present in a biological sample. Referencing the example illustrated in FIG. 30, mass spectrometry data sent to the search engines **3011** may have included four MS2 scan numbers (numbered 1 to 4 for simplicity) corresponding to four MS2 mass spectra that may each be associated with a particular type of glycomolecule. In this example, each of the search engines **3011** may have returned its own search results that predict the glycomolecules in an associated sample. In the illustrated example of FIG. 30, for Scan #2 corresponding to a particular mass spectrum, all search engines (Search Engines 1 to *i*) may

have predicted the same type of glycomolecule (which includes the glycan N(2)H(2)F(2)A(0)). In this example, the system may ensure that the consensus set includes this predicted glycomolecule search result for Scan #2.

[0046] In some cases, there may be a conflict among the different search engines, where two or more search engines disagree on the identity of a glycomolecule corresponding to a particular mass spectrum. In the illustrated example of FIG. 30, for Scan #1 corresponding to another mass spectrum, Search Engine 1 and Search Engine 3 may have predicted the same type of glycomolecule (which includes the glycan N(1)H(1)F(1)A(0)), while Search Engine *i* predicted a different type of glycomolecule (which includes the glycan N(1)H(1)F(1)A(0)). In such cases where there is a conflict among the search engines for a particular search result, the disclosed system may attempt to resolve such conflict by applying a conflict resolution rule set to determine what glycomolecule should correspond to the associated scan number in the consensus set, as provided herein. As such, in some embodiments, the initial consensus set may include all search results on which all the search engines agree, as well as the search results that the system determines following application of the system's conflict resolution rule set (in the cases where there are conflicts among the search engines). In some embodiments, as further described herein, a user may additionally or alternatively manually select among the different conflicting glycomolecules and may thus modify the initial consensus set. When the user is finished, this modified consensus set can be exported or otherwise used as a final consensus set. In some embodiments, the disclosed system may include an adaptive learning process (e.g., a suitable neural network process or other machine learning process) which may adapt the conflict resolution rule set in real-time based on the user's modifications to the consensus set. For example, if a user modifies the initial consensus set by substituting a particular search result for a different search result, the adaptive learning process may use that information to over time adapt the conflict resolution rule set based on detected patterns. In some embodiments, the modified consensus set may be exported or transferred into a separate adaptive machine learning process for a similar purpose.

[0047] In some embodiments, the user interface **3002** is further configured to display the unique identifier **3013** in association with the one or more representations. In some embodiments, the unique identifier **3013** is a mass spectrometry scan number (e.g., an MS2 scan number).

[0048] In some embodiments, the instructions, when executed by the processor, further causes the processor to provide the plurality of search result elements **3010a, b, …**

i, each search result element **3010** configured to recognize a user input to select one predicted glycomolecule among the identified search results **3014a** associated with the same unique identifier **3013** as the consensus search result **3024** for the unique identifier **3013**. In some embodiments, the user input comprises a touch and a scrolling activity, and wherein the user interface is further configured to scroll through the representations of the predicted glycomolecule search results **3014a** displayed in the search results element **3010** in response to the scrolling activity. In some embodiments, the user interface **3002** is configured to arrange the representations of the predicted glycomolecule search result **3014a** of each search results set **3012** with a respective search results element **3010** into an ordered list for each of the search results set **3012**. In some embodiments the ordered lists are ordered by the unique identifier **3013**. In some embodiments, the unique identifier **3013** is the mass spectrometry scan number as shown in the non-limiting examples of FIGS. 30, 31A-31B, 32, and 33. In some embodiments, each of the ordered lists is identically ordered. In some embodiments, the user interface **3002** is further configured to scroll through the ordered lists in a coordinated manner in response to a scrolling activity on one of the ordered lists by the user as shown in FIG. 31A.

[0049] FIG. 31A is a schematic diagram showing a non-limiting example, according to some embodiments of the disclosure, of dynamic interaction with a user interface presenting glycomolecule search results displayed in search result elements and a consensus set displayed in a consensus list element. In some embodiments, the system further includes a display **3001**. In some embodiments, the display includes a touchscreen. In some embodiments, the display **3001** further includes a sensor. In some embodiments, the sensor includes capacitive sensing to detect and measure proximity, pressure, position and displacement, force, humidity, fluid level, acceleration, or any combination thereof. In some embodiments, the display **3001** further includes an actuator. In some embodiments, the actuator provides haptic feedback. In some embodiments, the actuator provides haptic feedback in the form of a vibration. In some embodiments, the system further includes a human interface device. Exemplary human interface devices may include, but are not limited to, a keyboard, a mouse, a touchpad, the touchscreen, a digital pen, a finger of the user interfacing directly with a touchscreen, and/or the like. In some embodiments, as illustrated for example in FIG. 2A, the system or method recognizes a human interface device and the user interface **3002** may render a user interface element **3040**, which may be a cursor or similar element of any shape, that enables the user to interact with the user interface. In some embodiments, the user interface element **3040** may change its

appearance on the user interface **3002** dynamically depending upon the user input that is registered by the human interface device.

[0050] FIG. 31B is a schematic diagram showing a non-limiting example, according to some embodiments of the disclosure, of a user interface **3002** that presents a glycomolecule visualization. In some embodiments, the user interface is configured to recognize an input of a user, recognize a selected glycomolecule search results **3014a** when the input of the user hovers, clicks, taps, or otherwise interacts with a glycomolecule search results **3014a** shown in the search results elements **3010a, b, . . . *i*** or in the consensus list element **3020**, and present a glycomolecule visualization **3016** for the corresponding glycomolecule search results **3014a** as shown in FIG. 31B. In some embodiments, as illustrated in FIG. 31B, the glycomolecule visualization **3016** is configured to depict only the glycan portion of a selected glycomolecule. In some cases, this may be because it is generally more difficult to predict the glycan portions with a high degree of confidence (e.g., using mass spectrometry data), partly because glycans have a great degree of variability in structure (e.g., due to a large number of possible structure configurations). By contrast, the identities of conjugates (e.g., peptide, RNA, DNA, lipid) of glycans in a glycoconjugate may be relatively easy to predict. As such, users may find it of more interest to inspect the glycan portions of glycomolecules when trying to understand or validate a search result. However, it is contemplated that the glycomolecule visualization **3016** can be configured to depict an entire glycomolecule (in the case where the glycomolecule is a glyconjugate) or a portion thereof that includes more than just the glycan portion. For example, in the case of a glycopeptide, the glycomolecule visualization **3016** may depict the entire glycopeptide. In some embodiments, the input of the user includes, but is not limited to, a touch, a click, a press, a voice, or any combination thereof. In some embodiments, the glycomolecule visualization **3016** includes one or more potential structures of the selected glycomolecule search results **3014a**. Some other non-limiting examples of a glycomolecule visualization **3016** includes viewing the intact glycopeptide (i.e., the peptide backbone and all of its attached gylcans), glycolipids, glycoRNA, glycoDNA, and the like. In some embodiments, the glycomolecule visualization **3016** is shown in a modal window **3003** in the user interface **3002** as shown in FIG. 31B. In some embodiments, the glycomolecule visualization **3016** is shown in a popup window **3003** in the user interface **3002**, as in FIG. 31B. In some embodiments, the glycomolecule visualization **3016** is shown in a dedicated region **3003** of the user interface **3002** as shown in FIG. 31C.

[0051] Presenting the glycomolecule visualizations in the user interface **3002** may be advantageous in that such visualizations may provide additional context to the user. For example, these visualizations may be informative for a user engaged in manually curating the consensus list within the consensus element **3020**. In one non-limiting example, a user may gain additional insight from the glycomolecule visualization **3016** and may determine that a first glycomolecule suggested by a first search engine **3011a** is unlikely to be present in the particular sample associated with the respective search result, which results in the user to review the other glycomolecules identified from other search engines **3011b ... *i*** including the glycomolecule visualizations provided from these other search engines **3011b ... *i***. As explained herein, different glycomolecule search engines use their respective algorithms to make predictions about what glycomolecules are in a sample based on data (e.g., mass spectrometry data) associated with the sample. These algorithms are not perfect and sometimes make erroneous predictions. In some cases, users may be able to identify these erroneous predictions when the glycomolecules (or at least the glycan portion of the glycomolecules) are visualized. In some cases, a user may be able to make a subjective judgment or conclusion about a prediction of a particular glycomolecule after inspecting a potential structure of the glycomolecule. For example, the user may inspect a potential structure of a glycan portion of the glycomolecule as depicted in a visualization and determine that such a structure is unlikely to occur (or at least unlikely to be present in the sample). Currently, researchers using glycomolecule search platforms typically work with search results within spreadsheets or tables of text, making it difficult to make such judgments. In many cases, researchers may attempt to manually draw a structure or else input compositional information of individual glycans in separate software to visualize particular glycans of interest. By contrast, in some embodiments, the user interface **3002** disclosed here presents such visualizations within the user interface as the user is making such judgments when validating a search results set, or when curating a consensus list from multiple search results as further described herein. As such, the process of search results validation and the curation of a consensus list may be greatly enhanced by the disclosed user interface **3002**. Relatedly, the user interface **3002** in these embodiments also facilitates understanding of a search results set (or the associated sample) by providing immediate access to the depictions of the glycomolecules for side-by-side viewing as the user inspects the search results. The glycomolecule may be visualized using any suitable format. In some embodiments, glycans are visualized using Symbol Nomenclature for Glycans (SNFG), as illustrated in FIG. 31B.

[0052] In some embodiments, the user interface **3002** further includes notifications element 3030 to the user. The notifications element **3030** may be set up to inform the user of a glycomolecule that is identified in the consensus list element **3020** as having a confidence score **3015** above a default threshold or a user-set threshold. In some embodiments, the notifications element **3030** may additionally or alternatively be configured to display a subset of the search results from the consensus list that correspond to glycomolecules of particular characteristics, which may be set by the user. For example, the user may submit a user input specifying a criterion requesting a list of the glycomolecules that were determined by the search engines to be present at a high relative abundance (or within a specified range of abundances). In this example, search results (e.g., from the consensus list) that satisfy this criterion may be displayed within the notifications element **3030**. The user may specify any other suitable criteria, such as confidence score ranges, particular compositions (e.g., only glycomolecules with 2 fucoses), etc. In some embodiments, the notifications element **3030** further displays the criteria used by the conflict resolution rule set to populate the consensus list element **3020**. In one non-limiting example, with reference to FIG. 30, the notifications element **3030** may indicate to the user that the glycomolecule corresponding to Scan #1 on the consensus list element **3020** was determined at least in part by the majority (or largest plurality) of Search Engines **3011**. In another non-limiting example, again with reference to FIG. 30, the notifications element **3030** may indicate to the user that the glycomolecule corresponding to Scan #4 on the consensus list element **3020** was determined at least in part by associated confidence scores (e.g., the associated confidence score of 99 of search result **3014*i***). Any number of different such factors may be displayed in the notifications element **3030**. This display, by presenting the rationale behind why the system recommended a particular search result in the initial consensus set, may inform the user in deciding whether or not to agree with the search result recommendation in the initial consensus set. For example, if the user determines that a particular glycomolecule corresponding to a search result in the initial consensus set is unlikely to be present in the associated sample, and finds that the rationale presented in the notifications element 3030 is unpersuasive or unjustified, the user may modify the consensus set. The opposite may occur if the user finds the rationale persuasive or justified. In some embodiments, the notifications element **3030** may include a link, a file link, a local system hyperlink, a local network hyperlink, a cloud hyperlink, or the like to various log data generated from the decision-making process by the system and the user during automated and manual curation activity respectively, which may include various metadata

generated from the searching and curation activity. In some embodiments notifications element **3030** may display error messages. In some embodiments, the display **3001** includes an actuator. In these embodiments, when a notification element **3030** is caused to appear on the user interface **3002**, the actuator is triggered to vibrate the display **3001** or any other suitable device so as to alert the user.

[0053] In some embodiments, the user interface **3002** is configured to recognize an input of a user and permit the user to manually edit one or more of the consensus search results **3024** displayed in the consensus list element **3020**. In some embodiments, the user interface **3002** is further configured to permit the user to drag and drop one or more of the predicted glycomolecule search results **3014a** from one or more of the glycomolecule search results sets **3012** into the consensus list element **3020**, and thus incorporate those predicted glycomolecule search results **3014a** into the consensus set. In other words, in some embodiments the user interface **3002** is constructed and configured for user interaction. In some embodiments, the user interface **3002** is configured to recognize an input of a user and permit the user to drag and drop one or more of the predicted glycomolecule search results **3014** from one or more of the glycomolecule search results sets **3012** into the consensus set. In some embodiments, the search results further comprise a confidence score **3015** for each of the glycomolecules in the conflict search results **3220**, and wherein the user interface **3002** is configured to recognize an input of a user and permit the user to apply a custom confidence score **3015** to any of the glycomolecules in the conflict search results **3220**. In some embodiments, the one or more programs further include instructions, when executed by the processor, that update a machine learning process with the custom confidence scores. The machine learning process is configured to output recommendations and predictions. In a non-limiting example, the user interface **3002** permits the user to rate their own confidence score **3015** for each glycomolecule presented during the manual curation steps, which can inform the machine learning process and thereby further update the machine learning process to increase its ability for future recommendations/predictions. The machine learning process may include adaptive machine learning for real-time collection of user-provided information (which includes, but is not limited to, custom confidence scores) to provide real-time recommendations and predictions to the user.

[0054] With reference to Fig. 34, non-limiting embodiments of a system (e.g., computer system) of the present disclosure are provided. The computer system can include a memory **3420**, e.g., non-transitory memory, and a processor **3410** (e.g., a hardware

processor, a CPU), as disclosed herein. The system can include a human interface device **3430** (e.g., a display device **3435** to provide a user interface configured to display one or more search results elements and a consensus list element **3020**, as disclosed herein) and/or one or more other devices (e.g., a keyboard, mouse, touchpad, touchscreen, microphone, etc.) that allow a user to interact with the user interface, and/or to input data and/or instructions, as disclosed herein.

[0055] In some embodiments, the system includes a network interface **3440**. The network interface can be configured to allow the processor to access the internet or a remote server, such as a cloud-based server **3470**. In some embodiments, the processor is configured to retrieve search results via the network interface **3440**. In some embodiments, the system includes a database **3450**, e.g., a local database to store and access data (such as usage data or historical data, as disclosed herein). In some embodiments, the system includes a communication bus for the processor to communicate with components of the system.

[0056] In some embodiments, the glycomolecules include glycans, glycoproteins, glycolipids, glycoRNAs, glycoDNAs, or any combination thereof. In some embodiments, the glycomolecules comprise glycans. In some embodiments, the instructions, when executed by the processor, further causes the processor to convert one or more of the glycomolecule search results to a common format as shown in FIG. 36. In some embodiments, the common format is a universal glycan format **3700**. The universal glycan format **3700** can include any suitable representation for different monosaccharides that are or may be present in a glycan. In some embodiments, the universal glycan format includes a single-letter representation for different monosaccharides. In some embodiments, the universal glycan format uses 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more characters to represent each of the different monosaccharides. In some embodiments, the glycan format (e.g., the universal glycan format, or any one or more of the platform-specific glycan formats) provides the amount of a monosaccharide that is present in a glycan. In some embodiments, the glycan format (e.g., the universal glycan format, any one or more of the platform-specific glycan formats) lists the number of each monosaccharide that is present in the glycan. In some embodiments, the universal glycan format lists the amount of a monosaccharide that is not present in the glycan, e.g., as "0". Inclusion of monosaccharides that are not present in the glycan representation is advantageous because it ensures uniformity in the format of the representation. That is, in the universal glycan format **3700**, glycan representations for each glycan include

information about all the same sugars of interest (whether present or absent). This is in contrast to conventional formats, where glycan representations omit the monosaccharides that are absent. This uniformity of the universal glycan format is readily evident in FIG. 36, which show a number of glycan representations in conventional platform-specific formats and their counterparts in the universal glycan format. The uniformity allows for better readability, and also simplifies comparison and analysis of multiple glycomolecules.

[0057] In some embodiments, the universal glycan format **3700** includes the amount of fucose and/or sialic acid to represent a glycan, even if there is no fucose and/or sialic acid present in the glycan. In some embodiments, the universal glycan format includes the amount of N-acetylneuraminic acid and/or N-glycolylneuraminic acid to represent a glycan, even if there is no N-acetylneuraminic acid and/or N-glycolylneuraminic acid present in the glycan. In some embodiments, the universal glycan format includes a first number *l* corresponding to an amount of hexose; a second number *m* corresponding to an amount of N-acetylhexosamine; a third number *n* corresponding to an amount of fucose; and a fourth number *o* corresponding to an amount of N-acetylneuraminic acid, and wherein *l*, *m*, *n*, *o* are each an integer ≥ 0. In some embodiments, each representation of a monosaccharide in the universal glycan format is associated with the amount of the monosaccharide in a glycan. In some embodiments, a glycan format (e.g., the universal glycan format **3700**, any one or more of the platform-specific glycan formats **3710**, **3720**, **3730**, **3740**) may provide glycan structural information. For example, referencing Fig. 36, the PGLYCO3 (Structure) format uses the sequence of letters to represent the relative position of associated monosaccharides and further uses parentheses to denote the branching structure of the glycan. In some embodiments, the glycan format includes, for a given glycan, glycan structural information. In some embodiments, at least one of the platform-specific glycan formats (e.g., **3710**, **3720**, **3730**, **3740**) includes glycan structural information. In some embodiments, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50, 60, 70, 80, 90, 100 or more, or a number in a range defined by any two of the preceding values, or the platform-specific glycan formats includes glycan structural information. In some embodiments, none of the platform-specific glycan formats includes glycan structural information. In some embodiments, the universal glycan format **3700** includes glycan structural information. In some embodiments, the universal glycan format **3700** does not include glycan structural information.

[0058] In some embodiments, the glycan structural information is, for a given glycan, a single glycan structural arrangement. In some embodiments, the glycan structural information is, for a given glycan, the most likely glycan structural arrangement. In some embodiments, the glycan structural information is, for a given glycan, a set of all possible glycan structural arrangements. In some embodiments, the glycan structural information is, for a given glycan, a set of all biologically consistent glycan structural arrangements (e.g., consistent with the type of structural arrangements known to be found in glycans from the same biological source).

[0059] In some embodiments, the glycomolecule search results from each of the glycomolecule search engines is in a platform-specific format. In some embodiments, the platform-specific format is a platform-specific glycan format. In some embodiments, the platform-specific glycan format (e.g., **3710**, **3720**, **3730**, **3740**) can be suitable for representing glycan data as an input to a computer-implemented glycomolecule analysis routine, such as, but not limited to, a glycomolecule search engine. In some embodiments, the platform-specific glycan format (e.g., **3710**, **3720**, **3730**, **3740**) is a glycan format used by or compatible with a glycomolecule search engine **3010**. In general, a glycan database can be an input to a glycomolecule search engine **3010** when glycan data of the glycan database are represented in the glycan format used by or compatible with the glycomolecule search engine. In general, a glycan database cannot be an input to a glycomolecule search engine when glycan data of the glycan database are represented in a glycan format that is not used by or is incompatible with the glycomolecule search engine. In some embodiments, the first platform-specific glycan format **3710** is compatible with (e.g., a glycan database in the first platform-specific glycan format can be an input **3760** to) a glycomolecule search engine **3750**. In some embodiments, the second platform-specific glycan format **3720** is not compatible with (e.g., a glycan database in the second platform-specific glycan format cannot be an input **3770** to) the glycomolecule search engine **3750** with which the first platform-specific glycan format is compatible. In some embodiments, only the first platform-specific glycan format **3710** is compatible with the glycomolecule search engine **3750**. In some embodiments, the second platform-specific glycan format **3720** is compatible with (e.g., a glycan database in the second platform-specific glycan format can be an input to) a glycomolecule search engine with which the first platform-specific glycan format is not compatible.

[0060] In some embodiments, the platform-specific glycan format is a glycan format used by or compatible with a particular glycomolecule search engine or a particular

subset of glycomolecule search engines that is available for public and/or commercial use. In some embodiments, the glycomolecule search engine is a glycopeptide search engine. In some embodiments, the platform-specific glycan format is a glycan format used by or compatible with a glycopeptide search engine, such as, but not limited to, BYONIC, PGLYCO3, MSFRAGGER-GLYCO, METAMORPHEUS (also known as O-PAIR), and STRUCGP.

[0061] In some embodiments, the universal glycan format **3700** is different from at least one of the platform-specific glycan formats. In some embodiments, the universal glycan format **3700** is different from any of the platform-specific glycan formats. In some embodiments, the universal glycan format **3700** is the same as one of the platform-specific glycan formats.

[0062] Also provided herein is a method for curating a consensus list of glycomolecules identified in a biological sample on a user interface is provided. The method includes, at a computer system shown in FIG. 34 including one or more processors **3410**, and memory **3420** storing one or more programs configured to be executed by the one or more processors **3410**, the one or more programs including instructions for retrieving a plurality of glycomolecule search results sets **3012** (with reference to Fig. 30) for a biological sample from a plurality of glycomolecule search engines **3011a, b, ... *i*** wherein each glycomolecule search results set **3012** comprises glycomolecule search results **3014** identifying one or more predicted glycomolecule search results **3014a** as being present, wherein each predicted glycomolecule search results **3014a** is associated with a unique identifier **3013**.

[0063] In some embodiments, with reference to Fig. 30, the one or more programs further include instructions for determining a consensus set **3022**, wherein the consensus set **3022** includes, for each unique identifier **3013**, a consensus search result **3024** identifying one of the predicted glycomolecules **3024a** associated with the unique identifier **3013** in the glycomolecule search results sets **3012**. In some embodiments, the one or more programs further include instructions for providing a user interface **3002** configured to display one or more search results elements **3010a, b, ... *i*** configured to display one or more representations of the one or more predicted glycomolecule search results **3014a** from each of the plurality of glycomolecule search results sets **3012**. In some embodiments, representations corresponding to the glycomolecule search results **3014a, b, ... *i*** associated with the same unique identifiers **3013** are displayed in each of the search result elements **3010a, b, ... *i***. In some embodiments, the user interface **3002** is further

configured to display a consensus list element **3020** configured to display the consensus set **3022**. In some embodiments, the user interface is further configured to display the unique identifier **3013** associated with the one or more representations. In some embodiments, the one or more programs further include instructions for any of the embodiments described above.

[0064] In some embodiments, a computer-readable medium comprising instructions thereon, which when executed by a processor causes the processor to perform the method of any of the embodiments described above. The computer-readable medium can be any suitable medium that can store instructions for performing a method of the present disclosure. Suitable media include, without limitation, a hard drive, a SSD hard drive, a CD-ROM drive, a DVD-ROM drive, a floppy disk, a tape, a flash memory stick or card.

IV. COMPUTING SYSTEMS

[0086] A glycan data management or glycomolecule analysis system of the present disclosure can include any suitable non-transitory memory for holding the instructions, which when executed, perform any of the methods disclosed herein. In some embodiments, the non-transitory memory is structured to include one or more of a read-only memory (ROM), flash memory, dynamic random access memory (DRAM) (such as synchronous DRAM (SDRAM), double data rate (DDR SDRAM), or DRAM (RDRAM), and so forth), static memory (e.g., flash memory, static random access memory (SRAM), and so forth), and a data storage device (e.g., hard disk, magnetic tape, any other magnetic medium, CD-ROM, any other optical medium, RAM, PROM, EPROM, FLASH-EPROM, any other memory chip or cartridge, and/or any other medium from which a processor or computer is adapted to read).

[0087] The glycan data management or glycomolecule analysis system can include any suitable processor configured to access the input glycan data and execute the instructions. In some embodiments, the processor is structured to include one or more general-purpose processing devices such as a microprocessor, central processing unit (CPU), and the like. In some embodiments, the processor includes a complex instruction set computing (CISC) microprocessor, reduced instruction set computing (RISC) microprocessor, very long instruction word (VLIW) microprocessor, or a processor

implementing other instruction sets or processors implementing a combination of instruction sets.

[0088] With reference to Fig. 7, non-limiting embodiments of a glycan data management system **700** are provided. The glycan data management system can include a non-transitory memory **720** and a processor **710** (e.g., a hardware processor, a CPU), as disclosed herein. In some embodiments, the glycan data management system includes a user interface **730**, e.g., to allow a user to input data and/or instructions (e.g., a keyboard, mouse, touchpad, touchscreen, microphone, etc.), and/or to allow a user to receive output (e.g., visual output, such as a screen or monitor; audio output, such as a speaker; etc.) from the system. The non-transitory memory **720** can include instructions for implementing the methods described herein with the processor **710**. In some embodiments, the glycan data management system includes a network interface **740**. The network interface can be configured to allow the processor to access the internet or a remote server, such as a cloud-based server **770**. In some embodiments, the processor is configured to access the input glycan data via the network interface **740**. In some embodiments, the glycan data management system includes a database **750**, e.g., a local database to store and access data. In some embodiments, the glycan data management system includes a communication bus **715** for the processor to communicate with components of the system.

[0089] In some embodiments, the input glycan data is from a glycan database stored in a data storage device (e.g., hard disk, magnetic tape, any other magnetic medium, CD-ROM, any other optical medium, RAM, PROM, EPROM, FLASH-EPROM, any other memory chip or cartridge, and/or any other medium from which a processor or computer is adapted to read). The input glycan data can be stored at any suitable location. In some embodiments, the input glycan data and/or the glycan database is stored on a server. With reference to Fig. 8, in some embodiments, the server is a local server **820**, e.g., on the same local area network (LAN) as the glycan data management system **810**. In some embodiments, one or more servers are located remotely over the internet **830**, across one or more server locations, e.g., **840**, **850**, **860**, each having one or more remote servers.

[0090] In some embodiments, the input glycan data and/or the glycan database is stored at a first remote server location **840**. In some embodiments, the first remote server location also hosts a first glycomolecule search engine, e.g., a first glycoprotein search engine, where the glycan format of the input glycan data and/or the glycan database is compatible with the first glycomolecule search engine. In some embodiments, a processor of the glycan data management system **810** is configured to access the input data from the

first remote server location **840**, and to execute instructions to convert the input glycan data from the first platform-specific glycan format, compatible with the first glycomolecule search engine at the first remote server location, via the universal glycan format, to a second platform-specific glycan format. In some embodiments, the second platform-specific glycan format is compatible with a second glycomolecule search engine at the second remote server location **850**. In some embodiments, a processor of the glycan data management system **810** is configured to access the input data from the first remote server location **840**, and to execute instructions to convert the input glycan data from the first platform-specific glycan format, compatible with the first glycomolecule search engine at the first remote server location, via the universal glycan format, to a third platform-specific glycan format. In some embodiments, the second platform-specific glycan format is compatible with a third glycomolecule search engine at the third remote server location **860**.

[0091] Any method of the present disclosure can be performed on any suitable system as disclosed herein. Provided herein is a glycan data management system for handling data associated with glycans, where the system includes a processor; and non-transitory memory comprising instructions, which when executed by the processor causes the processor to perform any one of the methods, e.g., computer-implemented methods, disclosed herein. Also provided herein is an electronic system for comparing glycomolecule search outputs from two or more glycomolecule search engines, where the system includes a processor; and non-transitory memory comprising instructions, which when executed by the processor causes the processor to perform any one of the methods, e.g., computer-implemented methods, disclosed herein.

[0106] With reference to Fig. 14, non-limiting embodiments of a glycomolecule data analysis system **1400** are provided. The glycomolecule data analysis system can include a non-transitory memory **1420** and a processor **1410** (e.g., a hardware processor, a CPU), as disclosed herein. In some embodiments, the glycomolecule data analysis system includes a user interface **1430**, e.g., to allow a user to input data and/or instructions (e.g., a keyboard, mouse, touchpad, touchscreen, microphone, etc.), and/or to allow a user to receive output (e.g., visual output, such as a screen or monitor; audio output, such as a speaker; etc.) from the system. In some embodiments, the glycomolecule data analysis system includes a network interface **1440**. The network interface can be configured to allow the processor to access the internet or a remote server, such as a cloud-based server **1470**. In some embodiments, the processor is configured to access the search output via the

network interface **1440**. In some embodiments, the glycomolecule data analysis system includes a database **1450**, e.g., a local database to store and access data. In some embodiments, the glycomolecule data analysis system includes a communication bus **1415** for the processor to communicate with components of the system.

[0107] In some embodiments, the search output is stored in a data storage device (e.g., hard disk, magnetic tape, any other magnetic medium, CD-ROM, any other optical medium, RAM, PROM, EPROM, FLASH-EPROM, any other memory chip or cartridge, and/or any other medium from which a processor or computer is adapted to read). The search output can be stored at any suitable location. In some embodiments, the search output is stored on a server. With reference to Fig. 15, in some embodiments, the server is a local server **1520**, e.g., on the same local area network (LAN) as the glycomolecule data analysis system **1510**. In some embodiments, one or more servers are located remotely over the internet **1530**, across one or more server locations, e.g., **1540**, **1550**, **1560**, each having one or more remote servers.

[0108] In some embodiments, a first search output is stored at a first remote server location **1540**. In some embodiments, the first remote server location also hosts a first search output, e.g., a search output from a first glycomolecule search engine, where the search output is in a format specific to the first glycomolecule search engine. In some embodiments, a processor of the glycomolecule data analysis system **1510** is configured to access the first search output from the first remote server location **1540**, and to execute instructions to convert the first search output from the first search engine-specific format to a common format. In some embodiments, a second search output is stored at the second remote server location **1550**. In some embodiments, a processor of the glycomolecule data analysis system **1510** is configured to access the second search output from the second remote server location **1550**, and to execute instructions to convert the second search output from the second search engine-specific format to the common format. In some embodiments, a third search output is stored at the third remote server location **1560**. In some embodiments, a processor of the glycomolecule data analysis system **1510** is configured to access the third search output from the third remote server location **1560**, and to execute instructions to convert the third search output from the third search engine-specific format to the common format.

[0081] An electronic system of the present disclosure can include any suitable non-transitory memory, e.g., configured as disclosed herein, or for holding instructions, which when executed, perform any of the methods disclosed herein. In some embodiments,

the non-transitory memory is structured to include one or more of a read-only memory (ROM), flash memory, dynamic random access memory (DRAM) (such as synchronous DRAM (SDRAM), double data rate (DDR SDRAM), or DRAM (RDRAM), and so forth), static memory (e.g., flash memory, static random access memory (SRAM), and so forth), and a data storage device (e.g., hard disk, magnetic tape, any other magnetic medium, CD-ROM, any other optical medium, RAM, PROM, EPROM, FLASH-EPROM, any other memory chip or cartridge, and/or any other medium from which a processor or computer is adapted to read).

[0082] The electronic system can include any suitable processor configured to receive the search output, and analyze the search output, as provided herein. In some embodiments, the processor is structured to include one or more general-purpose processing devices such as a microprocessor, central processing unit (CPU), and the like. In some embodiments, the processor includes a complex instruction set computing (CISC) microprocessor, reduced instruction set computing (RISC) microprocessor, very long instruction word (VLIW) microprocessor, or a processor implementing other instruction sets or processors implementing a combination of instruction sets.

[0083] With reference to Fig. 26, non-limiting embodiments of an electronic system **2600** are provided. The electronic system can include a non-transitory memory **2620** and a processor **2610** (e.g., a hardware processor, a CPU), as disclosed herein. In some embodiments, the electronic system includes a user interface **2630**, e.g., to allow a user to input data and/or instructions (e.g., a keyboard, mouse, touchpad, touchscreen, microphone, etc.), and/or to allow a user to receive output (e.g., visual output, such as a screen or monitor; audio output, such as a speaker; etc.) from the system. In some embodiments, the electronic system includes a network interface **2640**. The network interface can be configured to allow the processor to access the internet or a remote server, such as a cloud-based server **2670**. In some embodiments, the processor is configured to access the search output via the network interface **2640**. In some embodiments, the electronic system includes a database **2650**, e.g., a local database to store and access data. In some embodiments, the electronic system includes a communication bus **2615** for the processor to communicate with components of the system.

[0084] In some embodiments, the search output is stored in a data storage device (e.g., hard disk, magnetic tape, any other magnetic medium, CD-ROM, any other optical medium, RAM, PROM, EPROM, FLASH-EPROM, any other memory chip or cartridge, and/or any other medium from which a processor or computer is adapted to read).

The search output can be stored at any suitable location. In some embodiments, the search output is stored on a server. With reference to Fig. 27, in some embodiments, the server is a local server **2720**, e.g., on the same local area network (LAN) as the electronic system **2710**. In some embodiments, one or more servers are located remotely over the internet **2730**, across one or more server locations, e.g., **2740**, **2750**, **2760**, each having one or more remote servers.

[0085] In some embodiments, a first search output is stored at a first remote server location **2740**. In some embodiments, the first remote server location also hosts a first search output, e.g., a search output from a first glycomolecule search engine. In some embodiments, a processor of the electronic system **2710** is configured to access the first search output from the first remote server location **2740**, and to execute instructions to analyze the search output as provided herein (e.g., determine that the first search output identifies a first set of glycomolecules as being of a first type and a second set of glycomolecules as being of a second type). In some embodiments, a second search output is stored at the second remote server location **2750**. In some embodiments, a processor of the electronic system **2710** is configured to access the second search output from the second remote server location **2750**, and to analyze the search output as provided herein (e.g., determine that the second search output identifies a third set of glycomolecules as being of the first type and a fourth set of glycomolecules as being of the second type). In some embodiments, a third search output is stored at the third remote server location **2760**. In some embodiments, a processor of the electronic system **2710** is configured to access the third search output from the third remote server location **2760**, and to execute instructions to analyze the search output as provided herein.

[0086] Any method of the present disclosure can be performed on any suitable system as disclosed herein. Provided herein is an electronic system that includes: a processor; and non-transitory memory comprising instructions, which when executed by the processor causes the processor to perform any method of the present disclosure (e.g., a method of identifying glycomolecules in a biological sample, as disclosed herein; a method of determining a glycomolecule profile of a biological sample, as disclosed herein).

[0109] Any method of the present disclosure can be performed on any suitable system as disclosed herein. Provided herein is an electronic system for comparing glycomolecule search outputs from two or more glycomolecule search engines, where the system includes a processor; and non-transitory memory comprising instructions, which

when executed by the processor causes the processor to perform any one of the methods, e.g., computer-implemented methods, disclosed herein.

Additional embodiments

[0080] The systems and methods of the present disclosure are further provided in the following non-limiting embodiments. In some embodiments, present systems and methods facilitate curation of a consensus list of identified glycomolecules from a plurality of glycomolecule search engines. In some embodiments, a system is a medical diagnostic or life sciences research method to identify the glycomolecule in a sample. In some embodiments, the method includes 5, 10, 15, 20, 25, 30, 40, 50 or more steps, or a number of steps in a range defined by any two of the preceding numbers. In some embodiments, at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 40, 50 or more steps, or a number of steps in a range defined by any two of the preceding numbers, of the method are designed to be performed in sequence. In some embodiments, at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 40, 50 or more steps, or a number of steps in a range defined by any two of the preceding numbers, of the method are designed to be performed in parallel.

[0081] In some embodiments, a system of the present disclosure is configured to provide notifications to the user of particular tasks that are time consuming or tasks that may create bottlenecks. In some embodiments, a system of the present disclosure is configured to suggest alternative tasks, or suggest abandoning a particular step. In some embodiments, a method of the present disclosure includes providing notifications to the user of particular tasks that are time consuming or tasks that may create bottlenecks. In some embodiments, a method of the present disclosure includes suggesting alternative tasks, or abandoning a particular step.

[0082] In any system or method of the present disclosure, in some embodiments, machine learning may be used for the predictions/notifications.

Additional system embodiments

[0092] In some embodiments, any of the systems of the present disclosure may be implemented in a cloud computing environment or cloud-based network. User interaction with the databases may be mediated via a central hub that stores and controls access to various interactions with the data. In some embodiments, the cloud computing environment may also provide sharing of protocols, analysis methods, libraries, as well as distributed processing for performing the analysis, and generating output reports. In some

embodiments, the system may be implemented in a computer browser, on-demand or on-line.

[0093] In some embodiments, instructions or software written to perform the systems or methods as described herein is stored in some form of computer readable medium, such as memory (e.g., non-transitory memory), CD-ROM, DVD-ROM, memory stick, flash drive, hard drive, SSD hard drive, server, mainframe storage system and the like.

[0094] In some embodiments, systems and methods may be written in any of various suitable programming languages, for example compiled languages such as C, C#, C++, Fortran, and Java. Other programming languages could be script languages, such as Perl, MatLab, SAS, SPSS, Python, Ruby, Pascal, Delphi, R and PHP. In some embodiments, the system or method is written in C, C#, C++, Fortran, Java, Perl, R, or Python. In some embodiments, the system may include an independent application with data input and data display modules. Alternatively, the system may include a computer software product and may include classes wherein distributed objects comprise applications including computational methods as described herein.

[0095] An assay instrument, desktop computer, laptop computer, or server which may contain a processor in operational communication with accessible memory may comprise the instructions for implementation of the systems and/or methods of the present disclosure. In some embodiments, a desktop computer or a laptop computer is in operational communication with one or more computer readable storage media or devices and/or outputting devices. An assay instrument, desktop computer and a laptop computer may operate under a number of different computer-based operational languages, such as those utilized by Apple® based computer systems or PC-based computer systems. An assay instrument, desktop and/or laptop computers and/or server system may further provide a computer interface for creating or modifying experimental definitions and/or conditions, viewing data results and monitoring experimental progress. In some embodiments, an outputting device may be a graphic user interface such as a computer monitor or a computer screen, a printer, a hand-held device such as a personal digital assistant (i.e., PDA, smartphone), a tablet computer, a hard drive, a server, a memory stick, a flash drive and the like.

[0096] A computer readable storage device or medium may be any device such as a server, a mainframe, a supercomputer, a magnetic tape system and the like. In some embodiments, a storage device may be located onsite in a location proximate to the assay

instrument, for example adjacent to or in close proximity to, an assay instrument. For example, a storage device may be located in the same room, in the same building, in an adjacent building, on the same floor in a building, on different floors in a building, etc., in relation to the assay instrument. In some embodiments, a storage device may be located off-site, or distal, to the assay instrument. For example, a storage device may be located in a different part of a city, in a different city, in a different state, in a different country, etc., relative to the assay instrument. In embodiments where a storage device is located distal to the assay instrument, communication between the assay instrument and one or more of a desktop, laptop, or server is typically via Internet connection, either wireless or by a network cable through an access point. In some embodiments, a storage device may be maintained and managed by the individual or entity directly associated with an assay instrument, whereas in other embodiments a storage device may be maintained and managed by a third party, typically at a distal location to the individual or entity associated with an assay instrument. In embodiments as described herein, an outputting device may be any device for visualizing data.

[0097] An assay instrument, desktop, laptop and/or server system may be used itself to store and/or retrieve computer implemented software programs incorporating computer code for performing and implementing computational methods as described herein, data for use in the implementation of the computational methods, and the like. One or more of an assay instrument, desktop, laptop and/or server may comprise one or more computer readable storage media for storing and/or retrieving software programs incorporating computer code for performing and implementing computational methods as described herein, data for use in the implementation of the computational methods, and the like. Computer readable storage media may include, but is not limited to, one or more of a hard drive, a SSD hard drive, a CD-ROM drive, a DVD-ROM drive, a floppy disk, a tape, a flash memory stick or card, and the like. Further, a network including the Internet may be the computer readable storage media. In some embodiments, computer readable storage media refers to computational resource storage accessible by a computer network via the Internet or a company network offered by a service provider rather than, for example, from a local desktop or laptop computer at a distal location to the assay instrument.

[0098] In some embodiments, computer readable storage media for storing and/or retrieving computer implemented software programs incorporating computer code for performing and implementing methods, e.g., computational methods, as described herein, data for use in the implementation of the methods, e.g., computational methods, and

the like, is operated and maintained by a service provider in operational communication with an assay instrument, desktop, laptop and/or server system via an Internet connection or network connection.

[0099] In some embodiments, a hardware platform for providing a computational environment comprises a processor (i.e., CPU) wherein processor time and memory layout such as random access memory (i.e., RAM) are systems considerations. For example, smaller computer systems offer inexpensive, fast processors and large memory and storage capabilities. In some embodiments, graphics processing units (GPUs) can be used. In some embodiments, hardware platforms for performing computational methods as described herein comprise one or more computer systems with one or more processors. In some embodiments, smaller computers are clustered together to yield a supercomputer network.

[0100] In some embodiments, methods, e.g., computational methods, as described herein are carried out on a collection of inter- or intra-connected computer systems (i.e., grid technology) which may run a variety of operating systems in a coordinated manner. For example, the CONDOR framework (University of Wisconsin-Madison) and systems available through United Devices are exemplary of the coordination of multiple stand-alone computer systems for the purpose of dealing with large amounts of data. These systems may offer Perl interfaces to submit, monitor and manage large sequence analysis jobs on a cluster in serial or parallel configurations.

EXAMPLES

Example 1

[0101] Human serum (SIGMA ALDRICH) was manually digested using trypsin (PROMEGA) and enriched for glycopeptides using the AssayMAP™ Bravo™ Platform with GlykoPrep™ cleanup cartridges (AGILENT TECHNOLOGIES). Samples were analyzed using an Orbitrap™ Exploris 480 mass spectrometer (THERMO FISHER SCIENTIFIC). An N-Linked Glycan database for individual search engines was built using the Glycan Database Converter to convert the search into the proper format for the BYONIC, PGLYCO3, and METAMORPHEUS formats. Identical searches were run on these search engines to yield respective glycopeptide search results. These search results from the various search engines were then converted into a universal format using a Glycomolecule Search Converter that is capable of converting search results into desired

formats. Confidence score comparisons were then compared by reverse tracing the scan numbers for individual experiments.

[0102] More specifically, we used the Glycan Database Converter to obtain PGLYCO3 glycan database input files for the BYONIC search engine's default N-glycan human plasma database. One human serum sample was enriched for glycopeptides using the AssayMAP Bravo Platform, analyzed on an Orbitrap™ Exploris 480 mass spectrometer coupled with Ultimate 3000 nanoLC system. Glycopeptide identification was achieved using searches on PGLYCO3 and BYONIC search engines with the same searching parameters. Our Glycomolecule Search Converter was later applied after search engine identifications to extract specific parameters including: MS2 Scan number, protein ID, peptide sequence, glycoform annotations, retention time, and confidence score output for each search engine.

[0103] For the human serum analyzed, PGLYCO3 identified 4498 glycopeptide-containing spectra from 120 glycoproteins, while BYONIC identified 8615 glycopeptide-containing spectra from 125 glycoproteins. The confidence scores reported by PGLYCO3 ranged from 5.03 to 164.36, while the confidence scores reported by BYONIC ranged from 0.04 to 1610.4. The magnitude of confidence score ranges were drastically different from PGLYCO3 compared to BYONIC. However, since the Glycomolecule Search Converter can effortlessly retrieve MS2 scan numbers and confidence scores, conversions between confidence scores for different search engines can be analyzed based on the same MS2 scan. With no additional confidence score filters, 81.5% (3670 out of 4498 PSMs) of the glycopeptide containing spectra reported by PGLYCO3 overlapped with BYONIC results. If filtered for the top 2000 PSMs based on the confidence scores of PGLYCO3 (which correlated to the top 44.5% of the glycopeptide containing spectra, the percentage of overlap increased to 91.5% (1870 out of 2000 PSMs). Interestingly, if we used the top 44.5% of glycopeptide-containing spectra that was ranked by BYONIC's confidence score, the percentage of overlap was about 76.6% (2936 out of 3834 PSMs), while a similar 81.5% overlap was observed when narrowing down to the top 35% of glycopeptide-containing spectra ranked by BYONIC's confidence score. In short, our glycopeptide converter enabled a direct comparison of the outputted results from PGLYCO3 and BYONIC while accounting for differences in the respective confidence scores between PGLYCO3 and BYONIC.

Example 2

[0104] This non-limiting example shows conversion of glycan representations between platform-specific glycan formats via a universal glycomolecular data format.

[0105] As shown in Fig. 9, glycan data with glycan representations in a universal glycan format ("Universal") is provided, where the sequential order of glycans was listed as the single letter representation of N, H, F, A, and G (from left to right) where each single letter representation had an associated integer value of 0 or greater corresponding to the amount of that glycan. The single letter representations N, H, F, A, and G correspond to HexNAc, Hex, fucose, NeuAc, and Neu5Gc, respectively. A BYONIC /Universal conversion rule set configured to convert each of the glycan representations in the Universal glycan format to and from a BYONIC glycan format ("BYONIC") is provided. The BYONIC glycan format is compatible with the BYONIC glycopeptide search engine. In this instance, the glycans HexNAc, Hex, NeuAc in the BYONIC format were converted to the single letter representation of N, H, and A, respectively, for the universal glycan format. Since one or more of the glycans fucose, NeuAc, and Neu5Gc were occasionally absent in the BYONIC format, the corresponding respective single letter representations were converted to F, A, and G, respectively, along with the zero integer to indicate the absent quantity.

[0106] A METAMORPHEUS/Universal conversion rule set configured to convert each of the glycan representations in the Universal glycan format to and from a METAMORPHEUS glycan format ("METAMORPHEUS") is provided. The METAMORPHEUS glycan format is compatible with the METAMORPHEUS glycopeptide search engine.

[0107] A PGLYCO (Composition)/Universal conversion rule set configured to convert each of the glycan representations in the Universal glycan format to and from a PGLYCO (Composition) glycan format that does not include glycan structural information ("PGLYCO (Composition)") is provided. The PGLYCO (Composition) glycan format is compatible with the PGLYCO glycopeptide search engine.

[0108] A PGLYCO (Structure)/Universal conversion rule set configured to convert each of the glycan representations in the Universal glycan format to and from a PGLYCO (Structure) glycan format that includes glycan structural information ("PGLYCO (Structure)") is provided. The PGLYCO (Structure) glycan format is compatible with the PGLYCO glycopeptide search engine.

[0109] As shown in Fig. 10, a database listing each glycan representation of the database in the corresponding PGLYCO (Structure), PGLYCO (Composition), Universal, BYONIC, and METAMORPHEUS glycan formats can be constructed.

[0110] For example, the BYONIC glycan format can be converted to the METAMORPHEUS glycan format by first converting the BYONIC glycan format to the Universal glycan format using the BYONIC /Universal conversion rule set, and then to the METAMORPHEUS glycan format using the METAMORPHEUS/Universal conversion rule set. A glycan database in the BYONIC glycan format can be converted to the METAMORPHEUS glycan format for use as an input to run the METAMORPHEUS glycopeptide search engine.

[0111] Similarly, a glycan database in any one of the platform-specific glycan formats can be converted to any one of the other platform-specific glycan formats via the Universal glycan format for use as an input to run the respective glycopeptide search engine compatible with the platform-specific glycan format.

Example 3

[0112] This non-limiting example shows conversion of a glycan representation from a format that does not include structural information, to a format that includes structural information.

[0113] A glycan database for searching a sample from human serum is obtained in a BYONIC-specific format. One of the glycans in the database is represented by its composition in the universal glycan format as: N(3)H(3)F(1)A(0). The universal glycan format does not include any structural information. A potential glycan structure for a glycan having the composition N(3)H(3)F(1)A(0) in a human serum sample is generated: (N(F)(N(H(H)(H(N))))).

[0114] Because the glycan database was for a human serum sample, glycan structures that would only be found in other sample types, e.g., human stool sample, were not included in the list of potential glycan structures.

Example 4

[0101] Human serum samples were manually digested using trypsin and enriched for glycopeptides using the AssayMAP™ Bravo™ Platform with GlykoPrep™ cleanup cartridges (AGILENT TECHNOLOGIES). Samples were analyzed using an Orbitrap™ Exploris 480 mass spectrometer (THERMO FISHER SCIENTIFIC). An N-

Linked Glycan database for three search engines (PGLYCO3, BYONIC, and METAMORPHEUS) was built using the Glycan Database Converter and identical searches were run on these search engines to yield respective glycopeptide search results. These search results from the various search engines were then converted into a universal format using a Glycopeptide Data Converter that is capable of converting search results into desired formats.

[0102] More specifically, we used the Glycan Database Converter on PGLYCO3 glycan database input files for 154 N-Linked glycans from the BYONIC search engine's N-glycan human plasma database. Glycopeptide identification was achieved using searches on PGLYCO3, BYONIC, and METAMORPHEUS search engines with the same search parameters. Our Glycopeptide Data Converter was applied after search engine identifications to extract specific parameters including: MS2 Scan number, peptide sequence, glycoform annotations, retention time, and confidence score output for each search engine.

[0103] For the human serum samples analyzed across five FAIMS collisional voltages, PGLYCO3 identified 3,038-6,801 glycopeptide-containing spectra (GCS), BYONIC identified 4,660-9,737 GCS, and METAMORPHEUS identified 3,140-6,818 GCS. However, since the Glycopeptide Data Converter can retrieve MS2 scan numbers and confidence scores, conversions between confidence scores for the different search engines can be analyzed based on the same MS2 scan. With no additional confidence score filters, 30.1% (3,266 of 10,868 PSMs) of glycopeptide containing spectra reported a shared space at FAIMS CV 45V by the PGLYCO3, BYONIC, and METAMORPHEUS search engines. As demonstrated by the relatively small degree of overlap for identified glycopeptides across the three search engines, a solution is needed to reconcile the differences and the Glycopeptide Data Converter enables a direct comparison of the outputted results from PGLYCO3, BYONIC, and METAMORPHEUS while accounting for differences in the respective confidence scores between the three different search engines.

Example 5

[0122] This non-limiting example shows conversion of search outputs from three glycopeptide search engines to a universal glycomolecular data format.

[0123] Search outputs from PGLYCO3, BYONIC, and METAMORPHEUS for the same experiment were obtained. The search results from PGLYCO3, BYONIC, and

METAMORPHEUS are in different formats, and cannot be easily compared. To facilitate comparison between the three search results, each search output was converted to a universal glycomolecular data format.

[0124] The PGLYCO3 search results were converted to the universal glycomolecular data format (Fig. 17). The universal glycomolecular data format included the identifier ("MS2#"), protein ID, peptide sequence, glycan data ("Glycan"), m/z value ("Precursor MZ"), charge number ("Z"), retention time ("RT"), confidence score ("Total score"), and the search engine. Parameters in the PGLYCO3 format that are not in the universal glycomolecular data format were removed. The peptide sequence from the PGLYCO3 format was converted to a common format by substituting "J" with "N". The glycan structural information in the PGLYCO format was not included in the universal glycomolecular data format.

[0125] The BYONIC search results were converted to the universal glycomolecular data format (Fig. 18). Parameters in the BYONIC format that are not in the universal glycomolecular data format were removed. The modification brackets and numbers in the peptide sequence of the BYONIC format were removed to convert to the universal glycomolecular data format. The amino acids flanking the peptide sequences provided in the BYONIC format were removed in the universal glycomolecular data format.

[0126] The METAMORPHEUS search results were converted to the universal glycomolecular data format (Fig. 19). Parameters in the METAMORPHEUS format that are not in the universal glycomolecular data format were removed. The format of the protein ID was the same between the METAMORPHEUS format and the universal glycomolecular data format.

[0127] Conversion of the search results from PGLYCO, BYONIC, and METAMORPHEUS from their respective search engine-specific formats to the universal glycomolecular data format facilitated comparison of the three search results (see bottom panels in Figs. 17, 18, 19).

Example 6

This non-limiting example shows conversion of specific parameters from a search engine-specific format to a universal glycomolecular data format.

[0129] Conversion of search results from PGLYCO3 to a universal glycopeptide data format involved conversion of protein ID, peptide sequence, retention time, and glycan

data from the PGLYCO3 format to the universal glycopeptide data format (Fig. 20). The protein identifier was converted to include only the portion required to uniquely identify the protein. The peptide sequence was converted to replace "J", indicating the asparagine residue to which the glycan is found to be attached, with "N". The retention time was converted from seconds to minutes. The glycan data was converted to a universal glycan format.

[0130] Conversion of search results from BYONIC to the universal glycopeptide data format involved conversion of the identifier, protein ID, peptide sequence, and glycan data from the BYONIC format to the universal glycopeptide data format (Fig. 21). The identifier was converted by removing extraneous information and leaving the identifier that is sufficient to identify the same analyte in the search results from other search engines. The protein identifier was converted to include only the portion required to uniquely identify the protein. The peptide sequence was converted to remove modification brackets and numbers, and to remove amino acids flanking the peptide sequence of the identified glycopeptide. The glycan data was converted to the universal glycan format.

[0131] Conversion of search results from METAMORPHEUS to the universal glycopeptide data format involved conversion of the m/z number (or precursor MZ), protein ID, and glycan data from the METAMORPHEUS format to the universal glycopeptide data format (Fig. 22). The m/z number (or precursor MZ) was converted by rounding the value to four decimal points. The protein identifier was converted typographically to the universal glycopeptide data format. The glycan data was converted to the universal glycan format.

Example 7

[0099] This non-limiting example shows identification of a conflicting subset of glycomolecules (e.g., glycopeptides) in search results from different glycomolecule search engines.

[0100] Search outputs from PGLYCO3 and BYONIC for the same glycopeptide mass spectrometry experiment were obtained. The search outputs were converted to a common format for direct comparison between each other. PGLYCO3 identifies a first set of glycopeptides as containing sialic acid (Fig. 28A, circle on the right), and BYONIC identifies a second set of glycopeptides as containing sialic acid (Fig. 28A, circle on the left). PGLYCO3 identifies a third set of glycopeptides as being fucosylated

(Fig. 28B, circle on the right), and BYONIC identifies a fourth set of glycopeptides as being fucosylated (Fig. 28B, circle on the left).

[0101] As illustrated in Figs. 28A and 28B, there is a conflicting subset of glycopeptides **610**. In this conflicting subset, a number of analytes are identified in the BYONIC search output as having sialic acid as illustrated in Fig. 28A, and the same analytes (e.g., as may be determined by their shared identifiers or "scan numbers") are identified in the PGLYCO3 search output as having fucose, thus creating a conflict. Analyte "X" may be one of the glycopeptides that is within this conflicting subset. In the illustrated example, Analyte X is identified as being a fucosylated glycopeptide in the PGLYCO3 search output (and not in the BYONIC search output) (Fig. 28B). The scan number for analyte "X" is used to determine that analyte "X" is identified as a glycopeptide containing sialic acid in the BYONIC search output (and not in the PGLYCO3 search output) (Fig. 28A). As disclosed herein, peptides with sialic acid and peptides with fucose are commonly confused for each other. This is due to the similarity in mass of some commonly occurring glycans, such as fucose and sialic acid, as illustrated in Fig. 24B (where the glycan is identical except that the first glycan includes two fucoses and the second glycan includes one N-acetylneuraminic acid (a sialic acid), as disclosed herein). Thus, the disclosed method may determine that there is a conflict.

[0102] The raw mass spectrometry data corresponding to analyte "X" with conflicting identification between the two search engines is pulled out of the sample data used to run the searches, and further analysis of the raw data is carried out to resolve the conflict. Based on further analysis, the analyte "X" is determined to be a fucosylated glycopeptide.

Example 8

[0103] This non-limiting example shows generating a glycopeptide profile based on a consensus list of overlapping sets of glycomolecules in search results from different glycomolecule search engines.

[0104] Search outputs from PGLYCO3 and BYONIC for the same glycomolecule mass spectrometry experiment were obtained. The search outputs were converted to a common format for direct comparison between each other. PGLYCO3 identifies a first set of glycomolecules with a PGLYCO3-specific confidence score equal to or above the PGLYCO3 threshold score of 50 (Fig. 29A, circle on the right), and BYONIC identifies a second set of glycomolecules with a BYONIC-specific confidence

score equal to or above the BYONIC threshold score of 200 (Fig. 29A, circle on the left). Glycomolecules identified in both the PGLYCO3 and BYONIC search results represents a first consensus list of glycomolecules in the sample from which the sample data used to run the searches were obtained.

[0105] Raising the threshold score for PGLYCO3 from 50 to 80 results in a third set of glycomolecules identified by PGLYCO3 (Fig. 29B, circle on the right) that is a subset of the first set of glycomolecules identified using the lower threshold score. The overlap between the third set of glycomolecules identified by PGLYCO3 using the higher threshold score and the second set of glycomolecules identified by BYONIC (Fig. 29B, circle on the left) identifies a second consensus list of glycomolecules in the sample from which the sample data used to run the searches were obtained.

[0106] The proportion of glycomolecules in the third set identified by PGLYCO3 that are in the second consensus list is larger than the proportion of glycomolecules in the first set identified by PGLYCO3 that are in the first consensus list. The proportion of glycomolecules in the second set identified by BYONIC that are in the second consensus list is larger than the proportion of glycomolecules in the second set identified by BYONIC that are in the first consensus list.

[0115] In at least some of the previously described embodiments, one or more elements used in an embodiment can interchangeably be used in another embodiment unless such a replacement is not technically feasible. It will be appreciated by those skilled in the art that various other omissions, additions, and modifications may be made to the methods and structures described above without departing from the scope of the claimed subject matter. All such modifications and changes are intended to fall within the scope of the subject matter, as defined by the appended claims.

[0116] It will be understood by those within the art that, in general, terms used herein, and especially in the appended claims (e.g., bodies of the appended claims) are generally intended as "open" terms (e.g., the term "including" should be interpreted as "including but not limited to," the term "having" should be interpreted as "having at least," the term "includes" should be interpreted as "includes but is not limited to," etc.). It will be further understood by those within the art that if a specific number of an introduced claim recitation is intended, such an intent will be explicitly recited in the claim, and in the absence of such recitation no such intent is present. For example, as an aid to understanding, the following appended claims may contain usage of the introductory phrases "at least one" and "one or more" to introduce claim recitations. However, the use of such phrases should not be construed to imply that the introduction of a claim recitation by the indefinite articles "a" or "an" limits any particular claim containing such introduced claim recitation to embodiments containing only one such recitation, even when the same claim includes the introductory phrases "one or more" or "at least one" and indefinite articles such as "a" or "an" (e.g., "a" and/or "an" should be interpreted to mean "at least one" or "one or more"); the same holds true for the use of definite articles used to introduce claim recitations. In addition, even if a specific number of an introduced claim recitation is explicitly recited, those skilled in the art will recognize that such recitation should be interpreted to mean at least the recited number (e.g., the bare recitation of "two recitations," without other modifiers, means at least two recitations, or two or more recitations). Furthermore, in those instances where a convention analogous to "at least one of A, B, and C, etc." is used, in general such a construction is intended in the sense one having skill in the art would understand the convention (e.g., " a system having at least one of A, B, and C" would include but not be limited to systems that have A alone, B alone, C alone, A and B together, A and C together, B and C together, and/or A, B, and C together, etc.). In those instances where a convention analogous to "at least one of A, B, or C, etc." is used, in general such

a construction is intended in the sense one having skill in the art would understand the convention (e.g., " a system having at least one of A, B, or C" would include but not be limited to systems that have A alone, B alone, C alone, A and B together, A and C together, B and C together, and/or A, B, and C together, etc.). It will be further understood by those within the art that virtually any disjunctive word and/or phrase presenting two or more alternative terms, whether in the description, claims, or drawings, should be understood to contemplate the possibilities of including one of the terms, either of the terms, or both terms. For example, the phrase "A or B" will be understood to include the possibilities of "A" or "B" or "A and B."

[0117] In addition, where features or aspects of the disclosure are described in terms of Markush groups, those skilled in the art will recognize that the disclosure is also thereby described in terms of any individual member or subgroup of members of the Markush group.

[0118] As will be understood by one skilled in the art, for any and all purposes, such as in terms of providing a written description, all ranges disclosed herein also encompass any and all possible sub-ranges and combinations of sub-ranges thereof. Any listed range can be easily recognized as sufficiently describing and enabling the same range being broken down into at least equal halves, thirds, quarters, fifths, tenths, etc. As a non-limiting example, each range discussed herein can be readily broken down into a lower third, middle third and upper third, etc. As will also be understood by one skilled in the art all language such as "up to," "at least," "greater than," "less than," and the like include the number recited and refer to ranges which can be subsequently broken down into sub-ranges as discussed above. Finally, as will be understood by one skilled in the art, a range includes each individual member. Thus, for example, a group having 1-3 articles refers to groups having 1, 2, or 3 articles. Similarly, a group having 1-5 articles refers to groups having 1, 2, 3, 4, or 5 articles, and so forth.

[0119] While various aspects and embodiments have been disclosed herein, other aspects and embodiments will be apparent to those skilled in the art. The various aspects and embodiments disclosed herein are for purposes of illustration and are not intended to be limiting, with the true scope and spirit being indicated by the following claims.

WHAT IS CLAIMED IS:

1. A glycan data management system for handling data associated with glycans, the system comprising:

a non-transitory memory comprising instructions for converting among a plurality of platform-specific glycan formats and a universal glycan format by applying at least one of a plurality of format conversion rule sets, the plurality of format conversion rule sets comprising:

a first rule set configured to convert glycan representations of a first platform-specific glycan format to and from the universal glycan format; and

a second rule set configured to convert glycan representations of a second platform-specific glycan format to and from the universal glycan format; and

a processor configured to:

access an input glycan data comprising one or more glycan representations in the first platform-specific glycan format, the second platform-specific glycan format, or the universal glycan format; and

execute the instructions to convert the input glycan data between:

the first platform-specific glycan format and the universal glycan format;

the second platform-specific glycan format and the universal glycan format; or

the first platform-specific glycan format and the second platform-specific glycan format via the universal glycan format.

2. The system of claim 1, wherein the processor is configured to execute the instructions to convert the input glycan data from the first platform-specific glycan format to the second platform-specific glycan format by applying: (1) the first rule set to convert the input glycan data from the first platform-specific glycan format to the universal glycan format; and (2) the second rule set to convert the input glycan data from the universal glycan format to the second platform-specific glycan format.

3. The system of claim 2, wherein the non-transitory memory further comprises instructions for obtaining the input glycan data from a first glycan database in the first platform-specific glycan format, and wherein execution of the instructions by the processor generates a second glycan database comprising glycan data in the second platform-specific glycan format.

4. The system of any one of the preceding claims, wherein the first platform-specific glycan format is compatible with a first glycomolecule search engine.

5. The system of claim 4, wherein the first glycomolecule search engine comprises a glycopeptide search engine.

6. The system of any one of the preceding claims, wherein the second platform-specific glycan format, but not the first platform-specific glycan format, is compatible with a second glycomolecule search engine.

7. The system of claim 5, wherein the universal glycan format is not compatible with the first or the second glycomolecule search engine.

8. The system of any one of the preceding claims, wherein the universal glycan format comprises compositional information only.

9. The system of any one of the preceding claims, wherein the universal glycan format comprises a single-letter representation for different monosaccharides.

10. The system of any one of the preceding claims, wherein the universal glycan format comprises a zero for a monosaccharide that is not present in the glycan data.

11. The system of any one of the preceding claims, wherein the universal glycan format comprises:

a first number *l* corresponding to an amount of hexose in the glycan data;

a second number *m* corresponding to an amount of N-acetylhexosamine in the glycan data;

a third number *n* corresponding to an amount of fucose in the glycan data; and

a fourth number *o* corresponding to an amount of N-acetylneuraminic acid in the glycan data,

wherein *l*, *m*, *n*, and *o* are each an integer 0 or greater.

12. The system of claim 11, wherein the universal glycan format further comprises a fifth number *p* corresponding to an amount of N-glycolylneuraminic acid, wherein *p* is an integer 0 or greater.

13. The system of any one of the preceding claims, wherein at least one of the first and second platform-specific glycan formats comprises glycan structural information.

14. A method of expanding the cross-compatibility of a glycan database, comprising:

providing to the glycan data management system of any one of claims 1-13 a glycan database comprising input glycan data comprising glycan representations in a first platform-specific glycan format; and

adding to the plurality of format conversion rule sets a third rule set configured to convert glycan representations of the third platform-specific glycan format to and from the universal glycan format, wherein none of the other rule sets among the plurality of format conversion rule sets is configured to convert glycan representations of the third platform-specific glycan format to and from the universal glycan format, wherein the third platform-specific glycan format is compatible with a third glycomolecule search engine and wherein the first platform-specific glycan format is not compatible with the third glycomolecule search engine.

15. The method of claim 14, further comprising causing the processor to access the input glycan data of the glycan database and executing the instructions to convert the input glycan data to the third platform-specific glycan format via the universal glycan format.

16. A method of converting a glycan data format, comprising:

receiving first glycan data comprising one or more glycan representations in a first platform-specific glycan format; and

converting the first glycan data from the first platform-specific glycan format to a second platform-specific glycan format via a universal glycan format, wherein the second platform-specific glycan format, but not the first platform-specific glycan format, is compatible with a first glycomolecule search engine.

17. The method of claim 16, wherein the first glycomolecule search engine comprises a first glycopeptide search engine.

18. The method of claim 16 or 17, wherein the first platform-specific glycan format is compatible with a second glycomolecule search engine.

19. The method of claim 16, wherein the second glycomolecule search engine comprises a second glycopeptide search engine.

20. The method of any one of claims 16-19, wherein converting the first glycan data comprises:

determining that the first glycan data comprises the one or more glycan representations in the first platform-specific glycan format; and

applying to the first glycan data: (1) a first rule set for converting the first platform-specific glycan format to the universal glycan format; and (2) a second rule set for converting the universal glycan data format to the second platform-specific glycan format.

21. The method of any one of claims 16-20, wherein the universal glycan format comprises compositional information only.

22. The method of any one of claims 16-21, wherein the universal glycan format comprises a single-letter representation for different monosaccharides.

23. The method of any one of claims 16-22, wherein the universal glycan format comprises a zero for a monosaccharide that is not present in the glycan data.

24. The method of any one of claims 16-23, wherein the universal glycan format comprises:

a first number *l* corresponding to an amount of hexose in the glycan data;

a second number *m* corresponding to an amount of N-acetylhexosamine in the glycan data;

a third number *n* corresponding to an amount of fucose in the glycan data; and

a fourth number *o* corresponding to an amount of N-acetylneuraminic acid in the glycan data,

wherein *l*, *m*, *n*, and *o* are each an integer 0 or greater.

25. The method of claim 24, wherein the universal glycan data format further comprises a fifth number *p* corresponding to an amount of N-glycolylneuraminic acid, wherein *p* is an integer 0 or greater.

26. The method of any one of claims 16-25, wherein at least one of the first and second platform-specific glycan formats comprises structural information.

27. A glycan data management system for handling data associated with glycans, the system comprising:

a processor; and

non-transitory memory comprising instructions, which when executed by the processor causes the processor to perform the method of any one of claims 16-26.

28. A method of converting a glycan format, comprising:

receiving glycan data comprising a glycan representation in a first format, the first format comprising compositional information; and

converting the glycan representation to a second format, the second format comprising compositional information,

wherein either:

(1) the first format further comprises glycan structural information, and the second format comprises compositional information only, and wherein converting the glycan representation comprises enumerating an amount of each of the monosaccharides in the glycan; or

(2) the first format comprises compositional information only, and the second format further comprises glycan structural information, and wherein converting the representation comprises:

receiving information about a source of the glycan data; and

generating a list of potential glycan structures based on the compositional information and the source of the glycan data.

29. The method of claim 28, wherein the source comprises human urine, human stool, human blood, or human serum.

30. The method of claim 28, wherein the source comprises a yeast, a plant, a non-human mammal, or a human.

31. A glycomolecule data analysis system for comparing glycomolecule search outputs from two or more glycomolecule search engines, the system comprising:

a processor; and

a non-transitory memory configured to cause the processor to:

access a first search output associated with a first glycomolecule search engine and a second search output associated with a second glycomolecule search engine, wherein the first search output comprises first glycomolecular data for each of one or more identified first glycomolecules in a first format, and wherein the second search output comprises second glycomolecular data for each of one or more identified second glycomolecules in a second format;

convert one or more of the first and second glycomolecular data to a common format, wherein the common format comprises the first format, the second format, or a universal glycomolecular data format; and

perform a comparison between the first and second glycomolecular data in the common format.

32. The system of claim 31, wherein one or more of the first and second glycomolecular data is converted to the common format by:

extracting one or more parameters associated with each of the one or more identified first and second glycomolecules in the first and second glycomolecular data, respectively, wherein the one or more parameters are in a search engine-specific format and comprises one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score; and

converting the one or more parameters associated with each of the one or more identified first and second glycomolecules to the common format.

33. The system of claim 31 or 32, wherein the common format is the universal glycomolecular data format.

34. The system of any one of the preceding claims, wherein the first and second search outputs correspond to the same glycomolecule mass spectrometry data run on the first and second glycomolecule search engines.

35. The system of any one of the preceding claims, wherein the non-transitory memory is configured to cause the processor to:

access a third search output associated with a third glycomolecule search engine, wherein the third search output comprises third glycomolecular data for each of one or more identified third glycomolecules in a third format;

convert one or more of the first, second, and third glycomolecular data to a common format, wherein the common format comprises the first format, the second format, the third format, or the universal glycomolecular data format; and

perform a comparison between any two or more of the first, second, and third glycomolecular data in the common format.

36. The system of claim 35, wherein the first, second, and third search outputs correspond to the same glycomolecule mass spectrometry data run on the first, second, and third glycomolecule search engines, respectively.

37. The system of claim 35 or 36, wherein the third glycomolecular data is converted to the universal glycomolecular data format by extracting one or more parameters associated with each of the one or more identified third glycomolecules in the third glycomolecular data, wherein the one or more parameters comprises one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score; and

converting the one or more parameters associated with each of the one or more identified third glycomolecules to the common format.

38. A glycomolecule data analysis system for comparing glycomolecule search outputs from two or more glycomolecule search engines, the system comprising:

a processor; and

a non-transitory memory configured to cause the processor to:

access two or more search outputs from two or more glycomolecule search engines, each search output (1) being associated with one of the two or more glycomolecule search

engines and (2) comprising glycomolecular data in a search engine-specific format for each of one or more glycomolecules identified by the glycomolecule search engine;

extract one or more parameters associated with each of the one or more identified glycomolecules in the glycomolecular data of each of the two or more search outputs, wherein the one or more parameters comprises an identifier; and

perform a comparison between the glycomolecular data from the two or more search outputs based on the one or more extracted parameters.

39. The system of claim 38, wherein the comparison is performed by identifying glycomolecules associated with the same identifier in the glycomolecular data from the two or more search outputs.

40. The system of claim 38 or 39, wherein the one or more parameters further comprises one or more of: glycan data, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score.

41. The system of any one of claims 38-40, wherein the two or more search outputs correspond to the same glycomolecule mass spectrometry data run on the two or more glycomolecule search engines, respectively.

42. The system of any one of claims 38-41, wherein the non-transitory memory is further configured to cause the processor to convert glycomolecular data of one or more of the two or more search outputs to a common format, wherein the common format comprises one of the search engine-specific formats or a universal glycomolecular data format.

43. The system of any one of claims 31-37 and 42, wherein the common format is the universal glycomolecular data format.

44. The system of any one of claims 31-37, 42 and 43, wherein the universal glycomolecular data format comprises identifying information for the glycomolecule search engine.

45. The system of any one of claims 31-37, and 42-44, wherein the universal glycomolecular data format comprises, for each identified glycomolecule, one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score.

46. The system of any one of the preceding claims, wherein the comparison comprises identifying an overlap between any two or more of the glycomolecular data.

47. The system of claim 46, wherein the overlap comprises a set of glycomolecules that are identified in any two or more of the glycomolecular data.

48. The system of claim 46 or 47, wherein the overlap comprises one or more of peptide sequences, protein IDs, retention times, m/z values, or charge numbers.

49. The system of any one of the preceding claims, wherein the comparison comprises identifying a difference between any two or more of the glycomolecular data.

50. The system of claim 49, wherein the difference comprises one or more sets of glycomolecules that are identified in one of the any two or more of the glycomolecular data as being different from the other one of the any two or more of the glycomolecular data.

51. The system of claim 49 or 50, wherein the difference comprises one or more of peptide sequences, protein IDs, retention times, m/z values, or charge numbers.

52. The system of any one of the preceding claims, wherein the glycomolecule is a glycopeptide, glycoprotein, glycolipid, glycoRNA, or glycoDNA.

53. The system of any one of the preceding claims, wherein the glycomolecule search engine is a glycopeptide search engine, glycoprotein search engine, glycolipid search engine, glycoRNA search engine, or glycoDNA search engine.

54. A method of comparing glycomolecule search outputs from two or more glycomolecule search engines, comprising:

receiving a first search output associated with a first glycomolecule search engine and a second search output associated with a second glycomolecule search engine, wherein the first search output comprises first glycomolecular data for each of one or more identified first glycomolecules in a first format, and wherein the second search output comprises second glycomolecular data for each of one or more identified second glycomolecules in a second format;

converting one or more of the first and second glycomolecular data to a common format, wherein the common format comprises the first format, the second format, or a universal glycomolecular data format; and

comparing the first and second glycomolecular data in the common format.

55. The method of claim 54, wherein the first and second glycomolecular data comprise one or more parameters associated with each of the one or more identified first and second glycomolecules in the first and second glycomolecular data, respectively, wherein the one or more parameters are in a search engine-specific format and comprises one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, and a confidence score, and wherein converting the one or

more of the first and second glycomolecular data to the common format comprises converting the one or more parameters to the common format.

56. The method of claim 54 or 55, wherein the first and second search outputs correspond to the same glycomolecule mass spectrometry data run on the first and second glycomolecule search engines.

57. The method of claim 56, further comprising providing a data set for running on each of the two or more glycomolecule search engines, wherein the data sets comprise the same glycomolecule mass spectrometry data.

58. The method of claim 57, further comprising:

generating the glycomolecule mass spectrometry data from a biological sample; and

providing the data set comprising the generated glycomolecule mass spectrometry data.

59. The method of claim 58, wherein the biological sample is from human urine, human stool, human blood, or human serum.

60. The method of any one of claims 54-59, wherein the two or more glycomolecule search engines are two or more glycopeptide search engines, the method further comprising providing a data set for running on each of the two or more glycopeptide search engines,

wherein the data sets comprise the same glycopeptide mass spectrometry data.

61. The method of claim 60, wherein each data set comprises a glycan database comprising glycan data in a glycan format that is compatible with a corresponding glycopeptide search engine of the two or more glycopeptide search engines.

62. The method of claim 61, wherein providing the data set further comprises:

providing the glycan database in a universal glycan format; and

converting the glycan database to a glycan format compatible with one of the two or more glycopeptide search engines.

63. The method of claim 61, wherein providing the data set further comprises:

receiving the glycan database in a glycan format compatible with one of the two or more glycopeptide search engines; and

converting the glycan database to the universal glycan format.

64. The method of any one of claims 54-63, comprising:

receiving a third search output associated with a third glycomolecule search engine, wherein the third search output comprises third glycomolecular data for each of one or more identified third glycomolecules in a third format;

converting one or more of the first, second, and third glycomolecular data to a common format, wherein the common format comprises the first format, the second format, the third format, or the universal glycomolecular data format; and

comparing any two or more of the first, second, and third glycomolecular data, each in the common format.

65. The method of claim 64, wherein the first, second, and third search outputs correspond to the same glycomolecule mass spectrometry data run on the first, second, and third glycomolecule search engines.

66. The method of any one of claims 54-65, wherein the common format is the universal glycomolecular data format.

67. The method of any one of claims 54-66, wherein the universal glycomolecular data format comprises identifying information for the glycomolecule search engine.

68. The method of any one of claims 54-67, wherein the universal glycopeptide data format comprises, for each identified glycomolecule, one or more of: glycan data, an identifier, a peptide sequence, a protein ID, a retention time, an m/z value, a charge number, or a confidence score.

69. The method of any one of claims 54-68, wherein the glycomolecule is a glycopeptide, glycoprotein, glycolipid, glycoRNA or glycoDNA.

70. The method of any one of claims 54-69, wherein the glycomolecule search engine is a glycopeptide search engine, glycoprotein search engine, glycolipid search engine, search engine or glycoDNA search engine.

71. A method for determining a glycomolecule profile of a biological sample, comprising:

providing to each of two or more glycomolecule search engines a data set comprising glycomolecule mass spectrometry data obtained from a biological sample, wherein the same glycomolecule mass spectrometry data are provided to each of the two or more glycomolecule search engines;

receiving an output from each of the two or more glycomolecule search engines run on the provided data set, wherein each output from the two or more glycomolecule search engines comprises glycomolecular data for each of one or more identified glycomolecules in a search engine-specific format;

converting glycomolecular data of one or more of the outputs from the two or more glycomolecule search engines to a common format, wherein the common format comprises one of the search engine-specific formats or a universal glycomolecular data format; and

determining a glycomolecule profile of the sample based on the glycomolecular data of the outputs from the two or more glycomolecule search engines in the common format.

72. An electronic system for comparing glycomolecule search outputs from two or more glycomolecule search engines, comprising:

a processor; and

non-transitory memory comprising instructions, which when executed by the processor causes the processor to perform the method of any one of claims 54-71.

73. A method of identifying glycomolecules in a biological sample, comprising:

receiving a first search output for a biological sample, the first search output being associated with a first glycomolecule search engine;

determining that the first search output identifies a first set of glycomolecules as being of a first type and a second set of glycomolecules as being of a second type;

receiving a second search output for the biological sample, the second search output being associated with a second glycomolecule search engine;

determining that the second search output identifies a third set of glycomolecules as being of the first type and a fourth set of glycomolecules as being of the second type; and

identifying among the first and second search outputs a conflicting subset of glycomolecules, wherein the conflicting subset of glycomolecules comprises:

(a) glycomolecules that are present in both the first set and the fourth set; or

(b) glycomolecules that are present in both the second set and the third set.

74. The method of claim 73, wherein the first set of glycomolecules is mutually exclusive of the third set of glycomolecules and the second set of glycomolecules is mutually exclusive of the fourth set of glycomolecules.

75. The method of any one of the preceding claims, wherein the first search output comprises:

a first set of identifiers associated with each glycomolecule of the first set of glycomolecules; and

a second set of identifiers associated with each glycomolecule of the second set of glycomolecules, and

wherein the second search output comprises:

a third set of identifiers associated with each glycomolecule of the third set of glycomolecules;

a fourth set of identifiers associated with each glycomolecule of the fourth set of glycomolecules, and

wherein identifying the conflicting subset of glycomolecules comprises:

(a) determining an identifier in the first set of identifiers that is the same as an identifier in the fourth set of identifiers; or

(b) determining an identifier in the second set of identifiers that is the same as an identifier in the third set of identifiers.

76. A method of identifying glycomolecules in a biological sample, comprising:

receiving a first search output for a biological sample, the first search output being associated with a first glycomolecule search engine;

receiving a second search output for the biological sample, the second search output being associated with a second glycomolecule search engine; and

determining that the first and second search outputs identify a conflicting subset of glycomolecules, wherein the conflicting subset of glycomolecules comprises one or more glycomolecules determined as being both of a first type in the first search output and of a second type in the second search output.

77. The method of claim 76, wherein the one or more glycomolecules of the conflicting subset of glycomolecules are determined as not being of the second type in the first search output.

78. The method of claim 76 or 77, wherein the one or more glycomolecules of the conflicting subset of glycomolecules are determined as not being of the first type in the second search output.

79. The method of any one of claims 76-78, wherein the first search output comprises a set of identifiers associated with each glycomolecule of a first set of

glycomolecules, wherein the second search output comprises a second set of identifiers associated with each glycomolecule of a second set of glycomolecules, and wherein determining that the first and second search outputs identify the conflicting subset of glycomolecules comprises:

identifying a first identifier in the first set of identifiers associated with a glycomolecule of the first type in the first set of glycomolecules; and

identifying a second identifier in the second set of identifiers associated with a glycomolecule of the second type in the second set of glycomolecules,

wherein the first and second identifiers are the same.

80. The method of any one of the preceding claims, further comprising generating a notification comprising a listing of the conflicting subset of glycomolecules.

81. The method of any one of the preceding claims, further comprising:

selecting a subset of data corresponding to the conflicting subset of glycomolecules; and

performing an additional analysis on the selected data to confirm an identity of the one or more glycomolecules of the conflicting subset of glycomolecules.

82. The method of any one of the preceding claims, wherein the first search output comprises a first set of confidence scores associated with the first or the second set of glycomolecules, and the second search output comprises a second set of confidence scores associated with the third or the fourth set of glycomolecules.

83. The method of claim 82, wherein each confidence score of the first set of confidence scores is equal to or greater than a first confidence score threshold, and each confidence scores of the second set of confidence scores is equal to or greater than a second confidence score threshold.

84. The method of claim 82 or 83, further comprising resolving the conflicting subset of glycomolecules based on a comparison of confidence scores of the first and second sets of confidence scores.

85. The method of claim 84, wherein resolving the conflicting subset of glycomolecules comprises, for each glycomolecule identified as being present in both the first set and the fourth set, or in both the second set and the third set, determining that the

glycomolecule is of the type identified by the search output comprising the higher confidence score associated with the glycomolecule.

86. The method of any one of claims 82-85, wherein the first and second sets of confidence scores are normalized.

87. The method of any one of the preceding claims, wherein glycomolecules of the first type and glycomolecules of the second type have masses within about 0-8 daltons of each other.

88. The method of any one of the preceding claims, wherein the first type comprises a sialic acid containing glycomolecule and the second type comprises a fucosylated glycomolecule.

89. The method of any one of the preceding claims, wherein the first search output is in a first format, wherein the second search output is in a second format, and wherein the method further comprises, converting one or both of the first and second search outputs to a common format, wherein the common format comprises the first format, the second format, or a universal glycomolecular data format.

90. A method of determining a glycomolecule profile of a biological sample, comprising:

receiving a first search output for a biological sample, the first search output being associated with a first glycomolecule search engine, wherein the first search output comprises a first set of glycomolecules associated with a first set of confidence scores;

receiving a second search output for the biological sample, the second search output being associated with a second glycomolecule search engine, wherein the second search output comprises a second set of glycomolecules associated with a second set of confidence scores; and

determining a glycomolecule profile of the biological sample based on (1) an overlap between the first and second search outputs and (2) the first and second set of confidence scores.

91. The method of claim 90, wherein the first search output comprises a first set of identifiers, and the second search output comprises a second set of identifiers, and wherein determining the glycomolecule profile further comprises:

determining the first set of identifiers associated with the first set of glycomolecules; and

determining the second set of identifiers associated with the second set of glycomolecules.

92. The method of claim 90 or 91, wherein the overlap comprises a set of glycomolecules in both the first and second search outputs, wherein each glycomolecule in the first set of glycomolecules is associated with a first confidence score of the first set of confidence scores that is equal to or greater than a first confidence score threshold, and each glycomolecule in the second set of glycomolecules is associated with a second confidence score of the second set of confidence scores that is equal to or greater than a second confidence score threshold.

93. The method of claim 92, further comprising setting the value of the first confidence score threshold and the value of the second confidence score threshold to optimize a percentage overlap between the search outputs among all the first and second glycomolecules having, respectively, first confidence scores equal to or greater than the first confidence score threshold and second confidence scores equal to or greater than the second confidence score threshold.

94. The method of claim 92 or 93, wherein determining the glycomolecule profile is based on an overlap between the first set of glycomolecules having first confidence scores within a first highest percentage and the second set of glycomolecules having second confidence scores within a second highest percentage.

95. The method of any one of claims 90-93, further comprising generating a consensus list of glycomolecules based on the overlap.

96. The method of claim 95, wherein generating the consensus list comprises applying a weighting factor to either one or both of the first and second sets of confidence scores.

97. The method of any one of claims 90-96, wherein the first search output is in a first format, and wherein the second search output is in a second format, wherein the method further comprises, before determining the glycomolecule profile, converting one or more of the first and second search outputs to a common format, wherein the common

format comprises the first format, the second format, or a universal glycomolecular data format.

98. The method of claim 97, wherein the common format is the universal glycomolecular data format.

99. The method of any one of claims 90-98, wherein the overlap comprises one or more of peptide sequences, protein IDs, retention times, m/z values, or charge states.

100. The method of any one of the preceding claims, wherein the first search output and second search output are generated by running the respective glycomolecule search engines on the same glycomolecule mass spectrometry data of the biological sample.

101. The method of any one of the preceding claims, wherein the glycomolecule is a glycopeptide, glycoprotein, glycolipid, glycoRNA, or glycoDNA.

102. The method of any one of the preceding claims, wherein the glycomolecule search engine is a glycopeptide search engine, glycoprotein search engine, glycolipid search engine, glycoRNA search engine, or glycoDNA search engine.

103. The method of any one of the preceding claims, wherein the biological sample is from human urine, human stool, human blood, or human serum.

104. An electronic system comprising:

a processor; and

non-transitory memory comprising instructions, which when executed by the processor causes the processor to perform the method of any one of the preceding claims.

105. A system for curating a consensus list of glycomolecules identified in a biological sample on a user interface, the system comprising:

a processor; and

a memory comprising instructions that, when executed by the processor, causes the processor to:

retrieve a plurality of glycomolecule search results sets for a biological sample from a plurality of glycomolecule search engines, wherein each glycomolecule search results set comprises glycomolecule search results

identifying one or more predicted glycomolecules as being present, wherein each predicted glycomolecule is associated with a unique identifier;

determine a consensus set comprising, for each unique identifier, a consensus search result identifying one of the predicted glycomolecules associated with the unique identifier in the glycomolecule search results sets;

provide a user interface configured to display:

one or more search results elements configured to display one or more representations corresponding to the one or more predicted glycomolecules from each of the plurality of glycomolecule search results sets; and

a consensus list element configured to display the consensus set.

106. The system of claim 104, wherein representations corresponding to the glycomolecule search results for predicted glycomolecules associated with the same unique identifiers are displayed in each of the search result elements.

107. The system of claim 105 or 106, wherein the user interface is further configured to display the unique identifier in association with the one or more representations.

108. The system of any one of claims 105-107, wherein the consensus set is determined at least in part by:

identifying conflicting search results and non-conflicting search results among the glycomolecule search results; and

populating the consensus list element with the non-conflicting search results.

109. The system of claim 108, wherein the glycomolecule search results are further evaluated by:

resolving the conflicting search results by applying a conflict resolution rule set to generate resolved conflicting search results; and

populating the consensus list element with the resolved conflicting search results.

110. The system of claim 109, wherein the conflict resolution rule set comprises identifying, for a particular conflicting search result associated with a particular unique

identifier, one or more plurality glycomolecules from the predicted glycomolecules of the conflicting search results associated with the particular unique identifier.

111. The system of claim 109, wherein the conflict resolution rule set comprises identifying, for a particular conflicting search result associated with a particular unique identifier, a majority glycomolecule from the predicted glycomolecules of the conflicting search results associated with the particular unique identifier.

112. The system of any one of claims 109-111, wherein the search results comprise a confidence score for each of the glycomolecules in the conflict search results, and wherein the conflict resolution rule set comprises identifying a consensus search result based on a comparison of the confidence scores of each search result among the conflict search results.

113. The system of any one of claims 105-112, wherein determining the consensus set comprises determining an initial consensus set, wherein the initial consensus set comprises a single predicted glycomolecule for each unique identifier.

114. The system of any one of claims 105-113, wherein the instructions, when executed by the processor, further causes the processor to provide the plurality of search result elements, each search result element configured to recognize a user input to select one predicted glycomolecule among the predicted glycomolecules associated with the same unique identifier as the consensus search result for the unique identifier.

115. The system of claim 114, wherein the user input comprises a touch and a scrolling activity, and wherein the user interface is further configured to scroll through the representations of the predicted glycomolecules displayed in the search results element in response to the scrolling activity.

116. The system of claim 114 or 115, wherein the user interface is configured to arrange the representations of the predicted glycomolecules of each search results set with a respective search results element into an ordered list for each of the search results set.

117. The system of claim 116, wherein each of the ordered lists is identically ordered.

118. The system of claim 116 or 117, wherein the user interface is further configured to scroll through the ordered lists in a coordinated manner in response to a scrolling activity on one of the ordered lists by the user.

119. The system of any one of claims 105-118, wherein the unique identifier is a mass spectrometry scan number, and

wherein the mass spectrometry scan number identifies a glycomolecule query in each of the glycomolecule search engines.

120. The system of any one of claims 105-119, further comprising a human interface device.

121. The system of any one of claims 105-120, further comprising a display.

122. The system of claim 121, wherein the display further comprises a sensor.

123. The system of claim 121 or 122, wherein the display further comprises an actuator.

124. The system of any one of claims 105-123, wherein the user interface is configured to:

recognize an input of a user;

recognize a selected glycomolecule when the input of the user hovers or interacts with the glycomolecule shown in the glycomolecule search results sets or the consensus set; and

present a visualized glycomolecule for the selected glycomolecule.

125. The system of claim 124, wherein the visualized glycomolecule comprises one or more potential structures of the selected glycomolecule.

126. The system of claim 124 or 125, wherein the visualized glycomolecule is shown in a modal window in the user interface.

127. The system of claim 124 or 125, wherein the visualized glycomolecule is shown in a popup window in the user interface.

128. The system of claim 124 or 125, wherein the visualized glycomolecule is shown in a dedicated region of the user interface.

129. The system of any one of claims 105-128, wherein the user interface is configured to:

recognize an input of a user and permit the user to manually edit one or more of the consensus search result; and optionally

permit the user to drag and drop one or more of the predicted glycomolecules from one or more of the glycomolecule search results sets into the consensus set.

130. The system of any one of claims 105-128, wherein the user interface is configured to:

recognize an input of a user and permit the user to drag and drop one or more of the predicted glycomolecules from one or more of the glycomolecule search results sets into the consensus set.

131. The system of any one of claims 105-130, wherein the glycomolecules comprise glycans, glycoproteins, glycolipids, glycoRNAs, glycoDNAs, or any combination thereof.

132. The system of any one of claims 105-131, wherein the glycomolecules comprise glycans.

133. The system of any one of claims 105-132, wherein the instructions, when executed by the processor, further causes the processor to convert one or more of the glycomolecule search results to a common format.

134. The system of claim 133, wherein the common format is a universal glycan format.

135. The system of claim 134, wherein the universal glycan format comprises:
a first number *l* corresponding to an amount of hexose;
a second number *m* corresponding to an amount of N-acetylhexosamine;
a third number *n* corresponding to an amount of fucose; and
a fourth number *o* corresponding to an amount of N-acetylneuraminic acid, and
wherein *l*, *m*, *n*, *o* are each an integer ≥ 0.

136. The system of any one of claims 105-135, wherein the glycomolecule search results from each of the glycomolecule search engines is in a platform-specific format.

137. A method for curating a consensus list of glycomolecules identified in a biological sample on a user interface, comprising:

at a computer system including one or more processors, and memory storing one or more programs configured to be executed by the one or more processors, the one or more programs including instructions for:

retrieving a plurality of glycomolecule search results sets for a biological sample from a plurality of glycomolecule search engines wherein each glycomolecule search results set comprises glycomolecule search results identifying one or more predicted glycomolecules as being present, wherein each predicted glycomolecule is associated with a unique identifier;

determining a consensus set, wherein the consensus set comprises, for each unique identifier, a consensus search result identifying one of the predicted glycomolecules associated with the unique identifier in the glycomolecule search results sets;

providing a user interface configured to display:

one or more search results elements configured to display one or more representations of the one or more predicted glycomolecules from each of the plurality of glycomolecule search results sets; and

a consensus list element configured to display the consensus set.

138. The method of claim 137, wherein representations corresponding to the glycomolecule search results for predicted glycomolecules associated with the same unique identifiers are displayed in each of the search result elements.

139. The method of claim 137 or 138, wherein the user interface is further configured to display the unique identifier associated with the one or more representations.

140. The method of any one of claims 137-139, wherein the determining the consensus list step comprises:

identifying conflicting search results and non-conflicting search results among the glycomolecule search results; and

populating the consensus list element with the non-conflicting search results.

141. The method of claim 140, the one or more programs further including instructions for:

resolving the conflicting search results to generate resolved conflicting search results; and

populating the consensus list element with the resolved conflicting search results.

142. The method of claim 141, the one or more programs further including instructions for:

applying a conflict resolution rule set, during the resolving the conflicting search results step, to resolve the conflicting search results,

wherein the search results further comprise a confidence score for each of the glycomolecules in the conflict search results.

143. The method of claim 142, wherein the conflict resolution rule set comprises identifying, for a particular conflicting search result associated with a particular unique identifier, one or more plurality glycomolecules from the predicted glycomolecules of the conflicting search results associated with the particular unique identifier.

144. The method of claim 142, wherein the conflict resolution rule set comprises identifying, for a particular conflicting search result associated with a particular unique identifier, a majority glycomolecule from the predicted glycomolecules of the conflicting search results associated with the particular unique identifier.

145. The method of claim 143 or 144, wherein the conflict resolution rule set further comprises identifying a consensus search result based on relative confidence scores of each search result among the conflict search results.

146. The method of claim 137, wherein the consensus set comprises an initial consensus set, and

wherein the initial consensus set comprises a single search predicted glycomolecule for each unique identifier.

147. The system of any one of claims 137-146, the one or more programs further including instructions for providing the plurality of search result elements, each of the search result elements configured to recognize a user input to select one predicted glycomolecule among the predicted glycomolecules associated with the glycomolecule search engine that is associated with the same unique identifier.

148. The system of any one of claims 137-147, the one or more programs further including instructions for arranging, in the user interface, the representations of the predicted glycomolecules of each search results set with a respective search results element into an ordered list for each of the search results set.

149. The method of claim 148, the one or more programs further including instructions for ordering each of the ordered lists identically.

150. The method of claim 147, wherein the input comprises a touch and a scrolling activity, and

wherein the user interface is further configured to scroll through the ordered lists at about a same scrolling speed and about a same scrolling direction in response to the scrolling activity on one of the ordered lists.

151. The method of any one of claims 137-150, wherein the unique identifier is a mass spectrometry scan number, and

wherein the mass spectrometry scan number identifies a glycomolecule query in each of the glycomolecule search engines.

152. The method of any one of claims 137-146, wherein the user interface is configured to:

recognize an input of a user and permit the user to manually edit one or more of the consensus search result; and optionally

permit the user to drag and drop one or more of the predicted glycomolecules from one or more of the glycomolecule search results sets into the consensus set.

153. The method of any one of claims 137-146, wherein the user interface is configured to:

recognize an input of a user and permit the user to drag and drop one or more of the predicted glycomolecules from one or more of the glycomolecule search results sets into the consensus set.

154. The method of claim 137, wherein the search results further comprise a confidence score for each of the glycomolecules in the conflict search results,

wherein the user interface is configured to:

recognize an input of a user and permit the user to apply a custom confidence score to any of the glycomolecules in the conflict search results.

155. The method of claim 154, the one or more programs further including instructions to update a machine learning process with the custom confidence scores,

wherein the machine learning process is configured to output recommendations and predictions.

156. The method of any one of claims 137-146, wherein the user interface is configured to:

recognize an input of a user;

recognize a selected glycomolecule when the input of the user hovers or interacts with the glycomolecule shown in the glycomolecule search results sets or the consensus set; and

present a visualized glycomolecule for the selected glycomolecule.

157. The method of any claim 156, wherein the visualized glycomolecule comprises one or more potential structures of the selected glycomolecule.

158. The method of claim 156 or 157, wherein the visualized glycomolecule is shown in a modal window in the user interface.

159. The method of claim 156 or 157, wherein the visualized glycomolecule is shown in a popup window in the user interface.

160. The method of claim 156 or 157, wherein the visualized glycomolecule is shown in a dedicated region of the user interface.

161. The method of any one of claims 147-160, wherein the glycomolecules comprise glycans, glycolipids, glycoRNAs, glycoDNAs, or any combination thereof.

162. The method of any one of claims 147-160, wherein the glycomolecules comprise glycans.

163. The method of claim 161 or 162, wherein the representation of the at least one glycomolecule comprises a glycan representation in a universal glycan format.

164. The method of claim 163, wherein the universal glycan format comprises:

a first number *l* corresponding to an amount of hexose;

a second number *m* corresponding to an amount of N-acetylhexosamine;

a third number *n* corresponding to an amount of fucose; and

a fourth number *o* corresponding to an amount of N-acetylneuraminic acid, and

wherein *l*, *m*, *n*, *o* are each an integer ≥ 0.

165. A computer-readable medium comprising instructions thereon, which when executed by a processor causes the processor to perform the method of any one of claims 137-164.

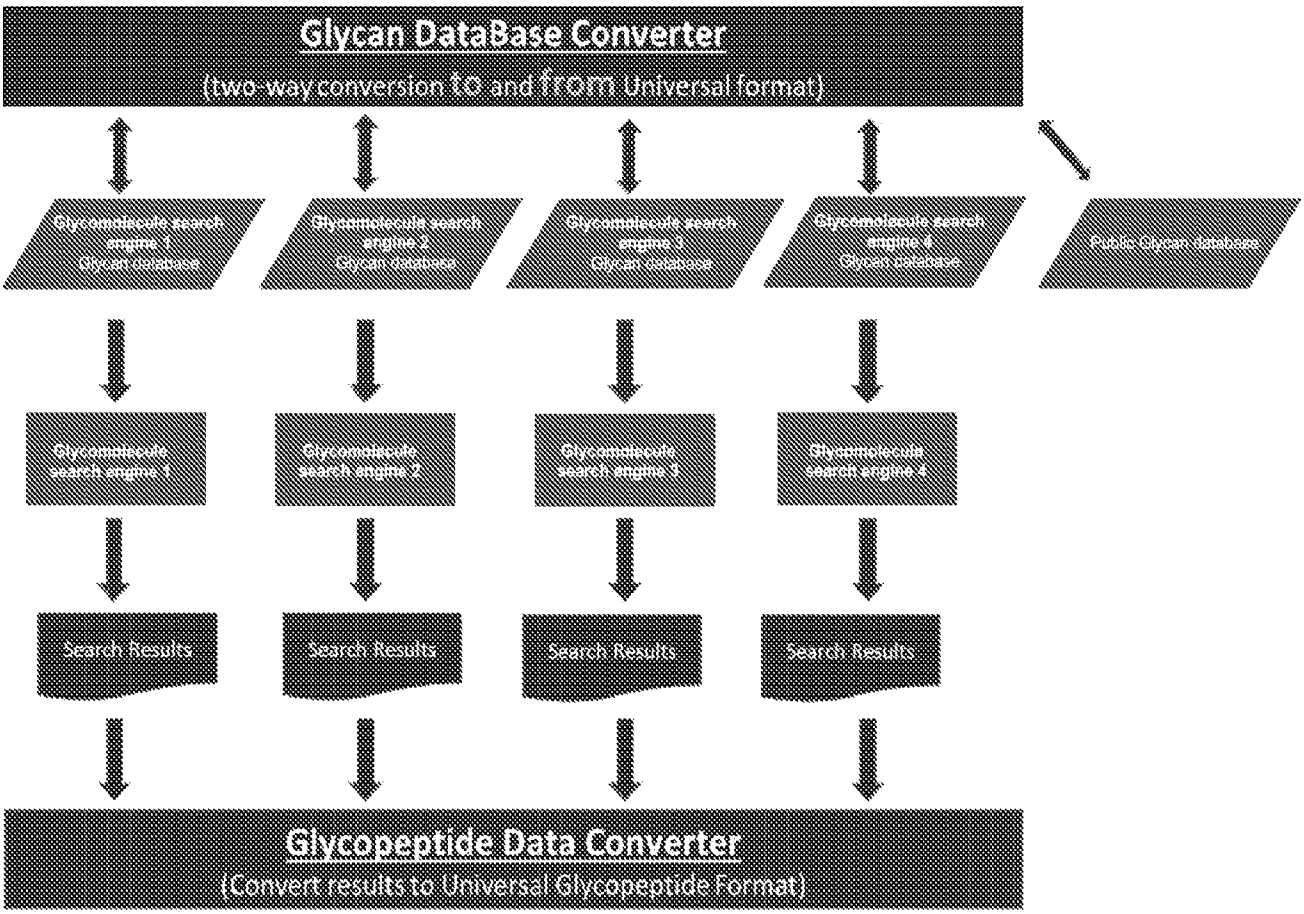

Glycan DataBase Converter
(two way conversion to and from Universal format)
Glycomolecule search engine 1 Glycan database
Glycomolecule search engine 2 Glycan database
Glycomolecule search engine 3 Glycan database
Glycomolecule search engine 4 Glycan database
Public Glycan database
Glycomolecule search engine 1
Glycomolecule search engine 2
Glycomolecule search engine 3
Glycomolecule search engine 4
Search Results
Search Results
Search Results
Search Results
Glycopeptide Data Converter
(Convert results to Universal Glycopeptide Format)

FIG. 1

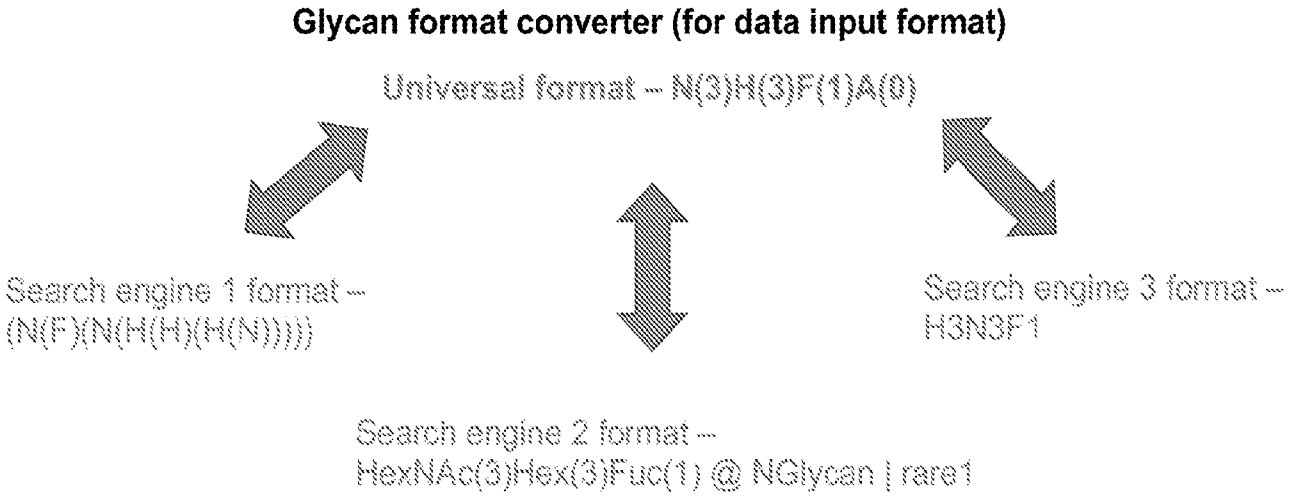

Glycan format converter (for data input format)
Universal format – N(3)H(3)F(1)A(0)
Search engine 1 format –
(N(F)(N(H(H)(H(N)))))
Search engine 3 format –
H3N3F1
Search engine 2 format –
HexNAc(3)Hex(3)Fuc(1) @ NGlycan | rare1

FIG. 2

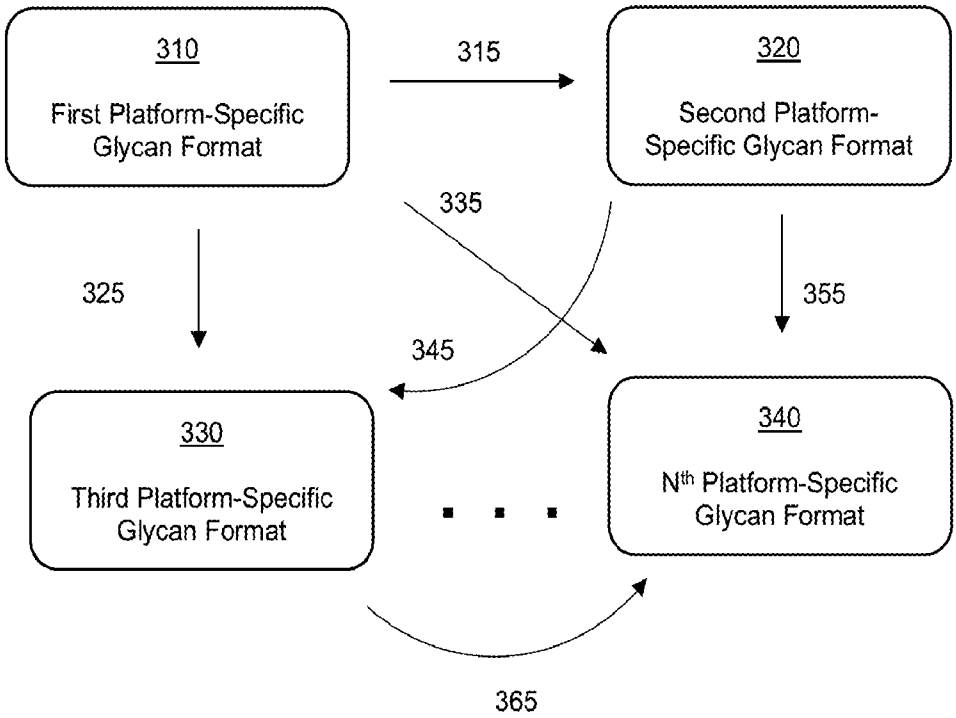

310
First Platform-Specific Glycan Format
315
320
Second Platform-Specific Glycan Format
335
325
355
345
330
Third Platform-Specific Glycan Format
340
Nth Platform-Specific Glycan Format
365

FIG. 3

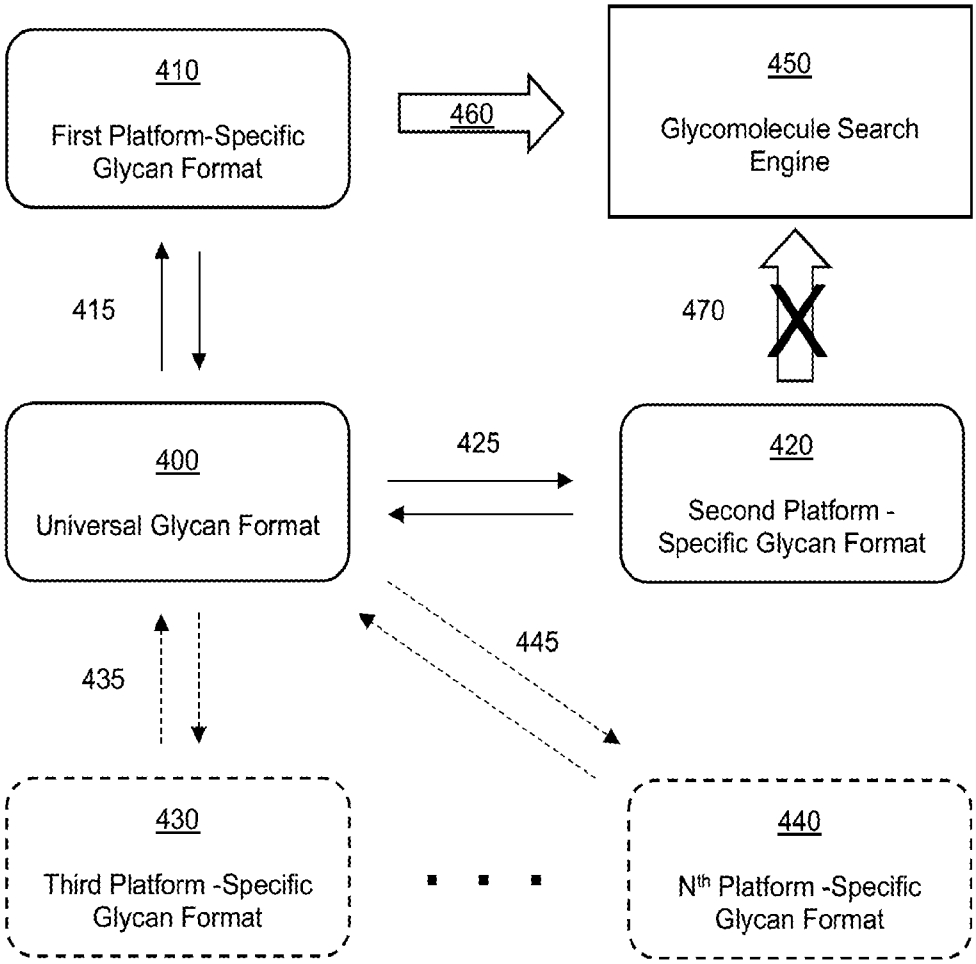

410
First Platform-Specific Glycan Format
460
450
Glycomolecule Search Engine
415
470
400
Universal Glycan Format
425
420
Second Platform - Specific Glycan Format
435
445
430
Third Platform -Specific Glycan Format
440
Nth Platform -Specific Glycan Format

FIG. 4

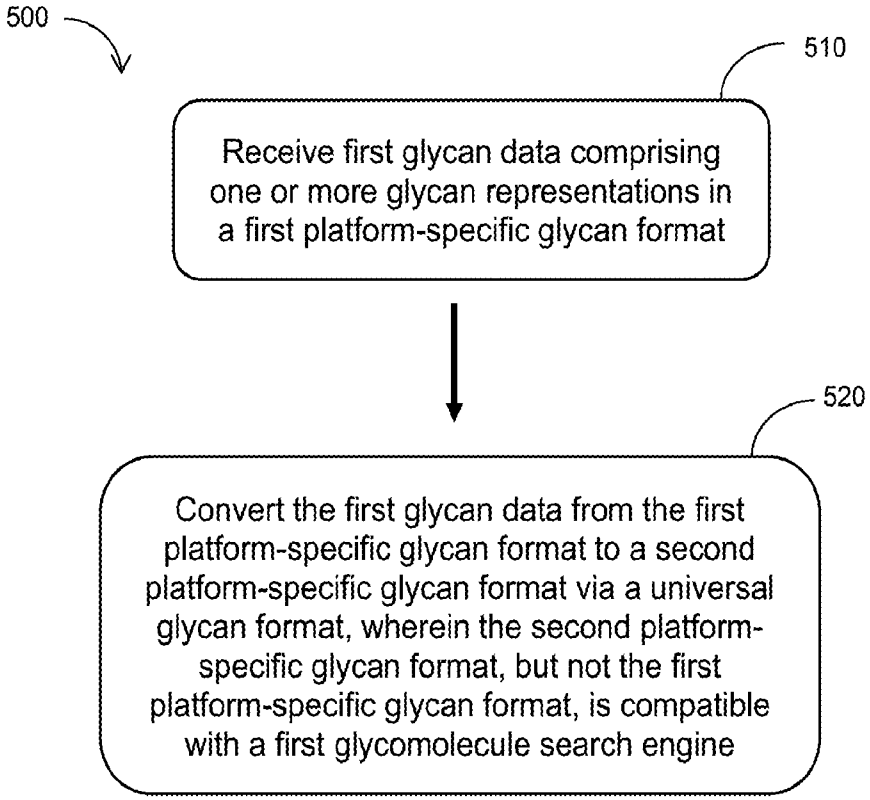

500
510
Receive first glycan data comprising one or more glycan representations in a first platform-specific glycan format
520
Convert the first glycan data from the first platform-specific glycan format to a second platform-specific glycan format via a universal glycan format, wherein the second platform-specific glycan format, but not the first platform-specific glycan format, is compatible with a first glycomolecule search engine

FIG. 5

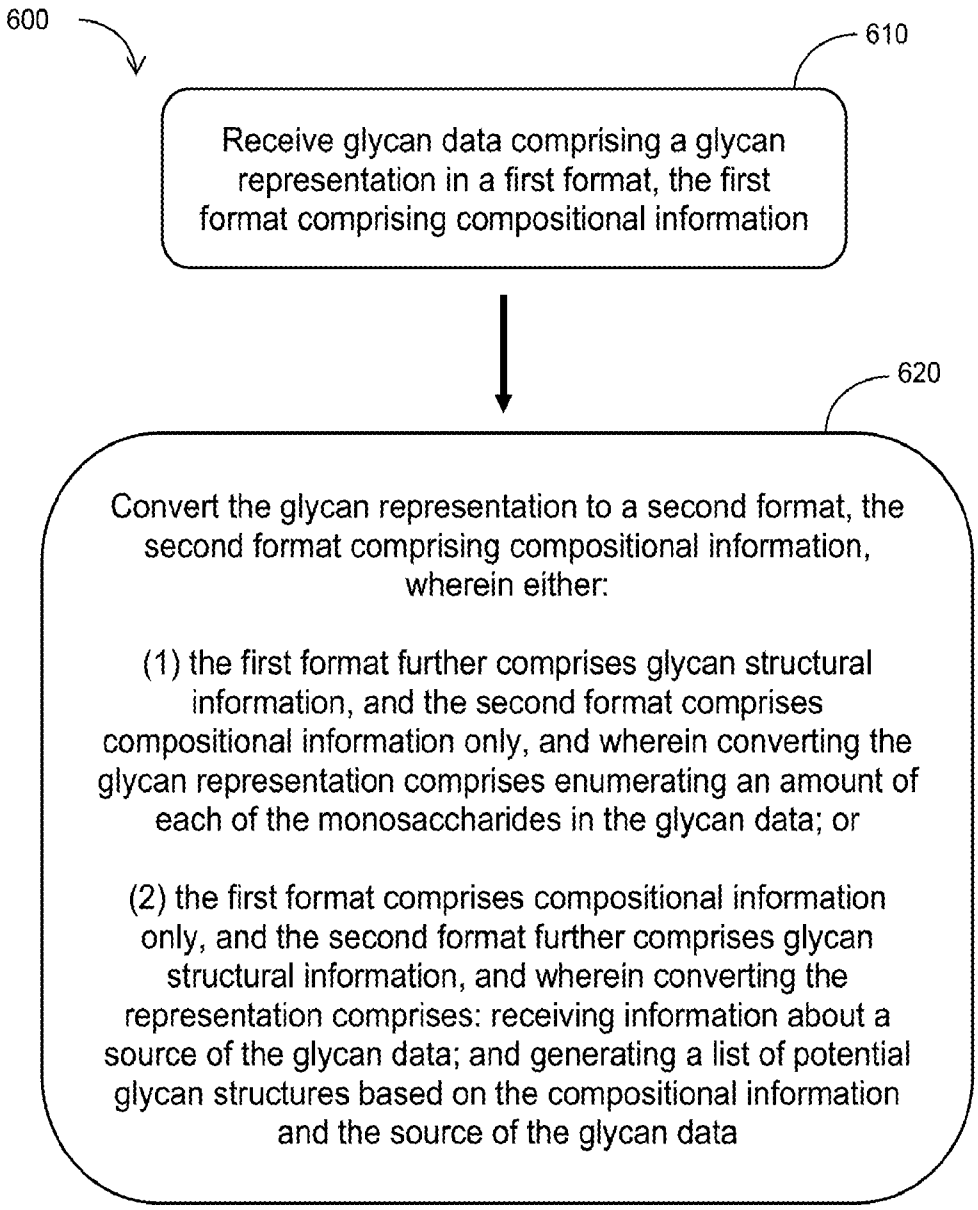

600
610
Receive glycan data comprising a glycan representation in a first format, the first format comprising compositional information
620
Convert the glycan representation to a second format, the second format comprising compositional information, wherein either:
(1) the first format further comprises glycan structural information, and the second format comprises compositional information only, and wherein converting the glycan representation comprises enumerating an amount of each of the monosaccharides in the glycan data; or
(2) the first format comprises compositional information only, and the second format further comprises glycan structural information, and wherein converting the representation comprises: receiving information about a source of the glycan data; and generating a list of potential glycan structures based on the compositional information and the source of the glycan data

FIG. 6

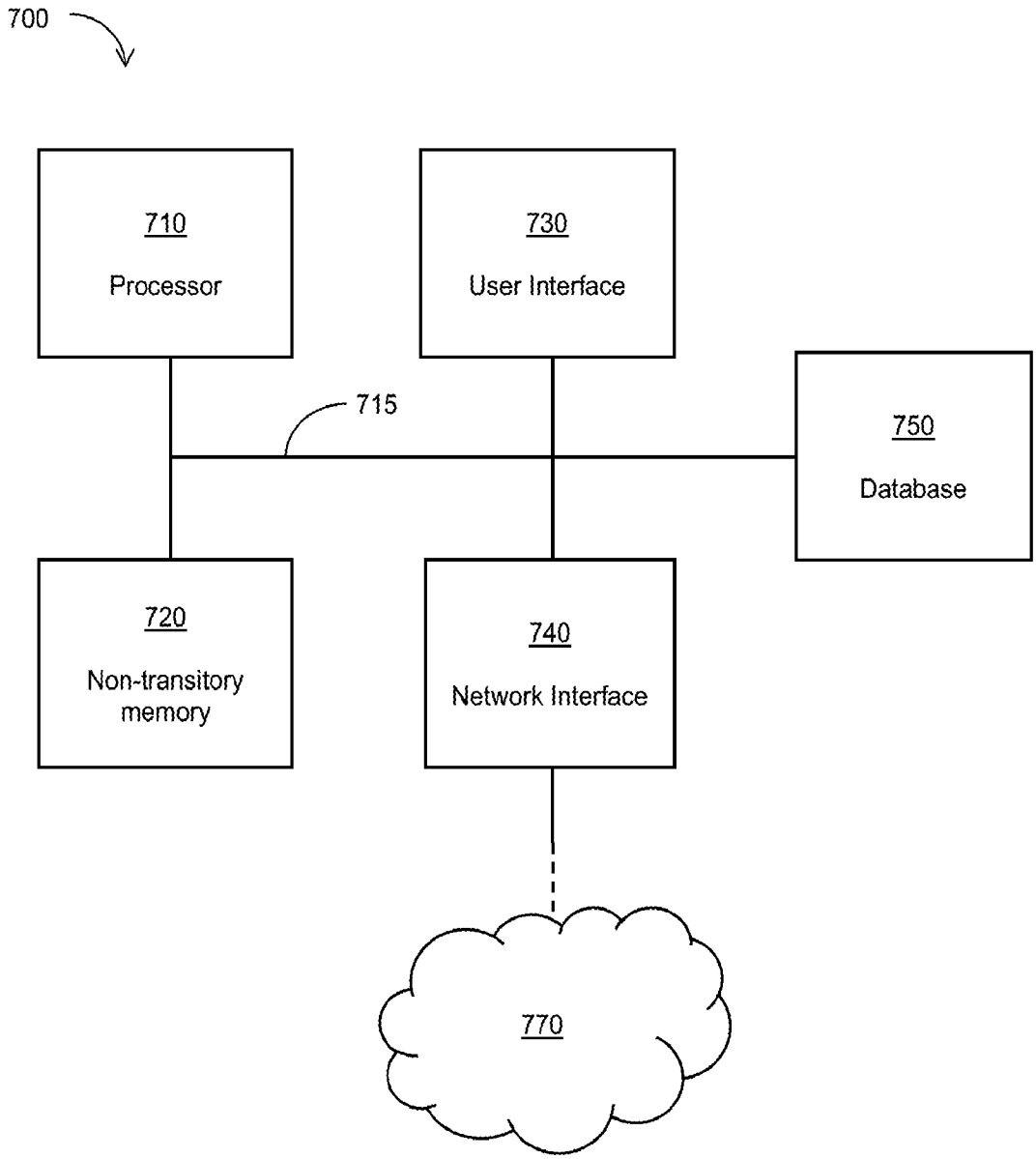

700
710
Processor
730
User Interface
715
750
Database
720
Non-transitory memory
740
Network Interface
770

FIG. 7

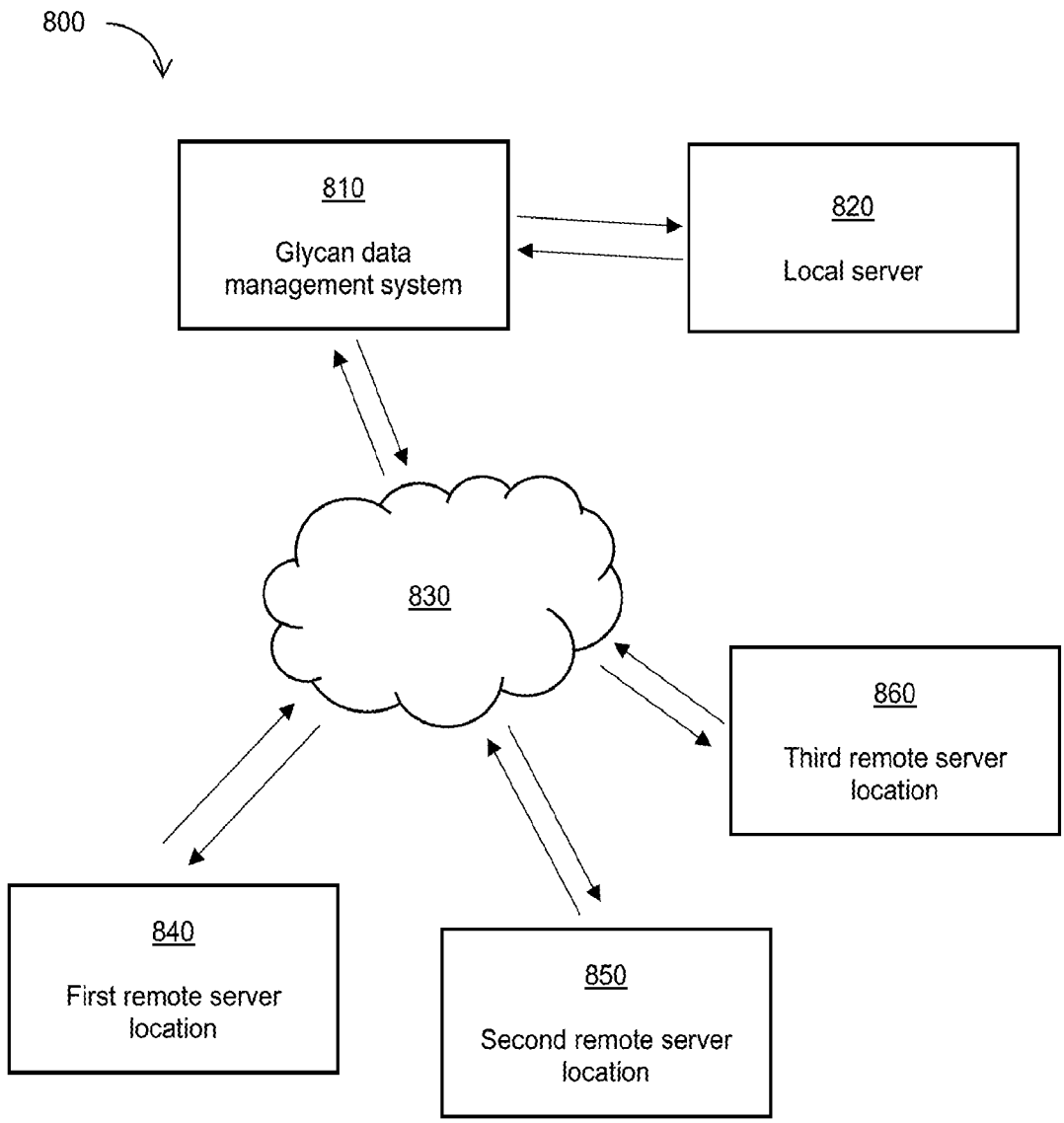

800
810
Glycan data management system
820
Local server
830
860
Third remote server location
840
First remote server location
850
Second remote server location

FIG. 8

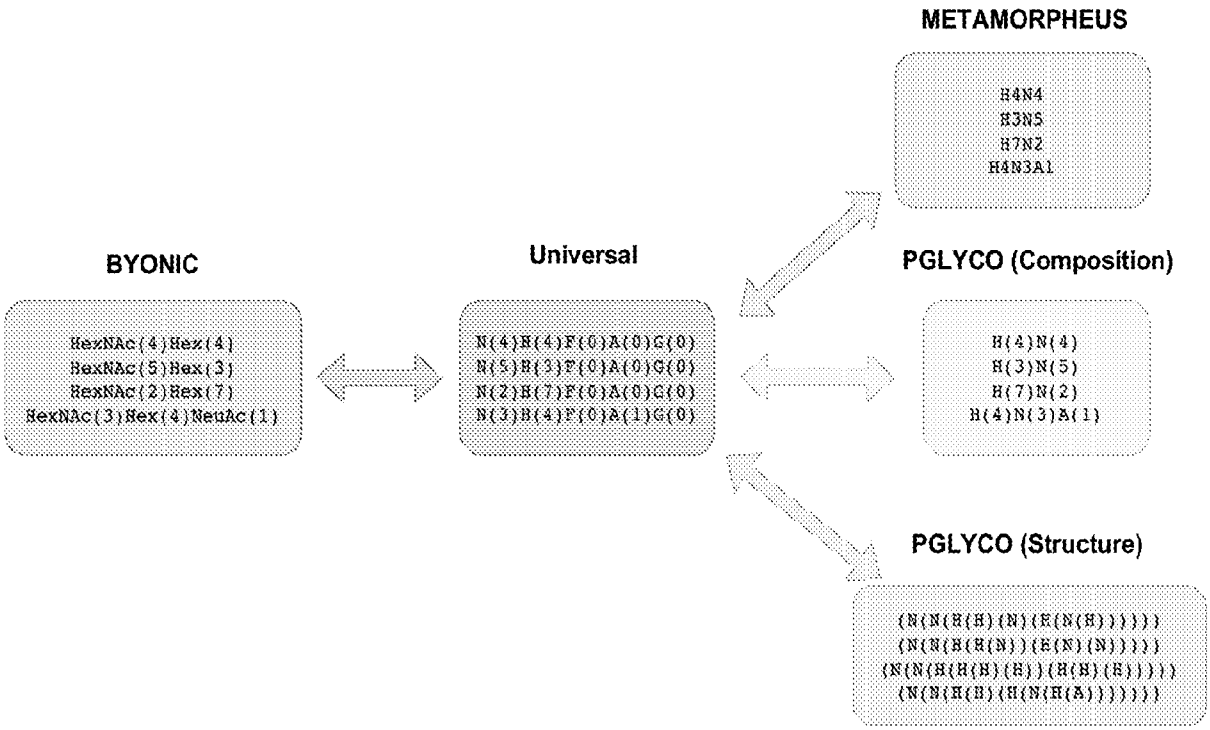

METAMORPHEUS
H4N4
H3N5
H7N2
H4N3A1
BYONIC
HexNAc(4)Hex(4)
HexNAc(5)Hex(3)
HexNAc(2)Hex(7)
HexNAc(3)Hex(4)NeuAc(1)
Universal
N(4)H(4)F(0)A(0)G(0)
N(5)H(3)F(0)A(0)G(0)
N(2)H(7)F(0)A(0)G(0)
N(3)H(4)F(0)A(1)G(0)
PGLYCO (Composition)
H(4)N(4)
H(3)N(5)
H(7)N(2)
H(4)N(3)A(1)
PGLYCO (Structure)
(N(N(H(H)(N)(H(N(H))))))
(N(N(H(H(N))(H(N)(N)))))
(N(N(H(H(H)(H))(H(H)(H)))))
(N(N(H(H)(H(N(H(A)))))))

FIG. 9

	A	B	C	D	E
1	PGLYCO3_structure	PGLYCO3_composition	Universal	BYONIC	METAMORPHEUS
2	{N(N{H(H){H}}}}	H{3}N{2}	N{2}H{3}F{0}A{0}G{0}	HexNAc{2}Hex{3}	H3N2
3	{N(F){N(H(H){H}}}}	H{3}N{2}F{1}	N{2}H{3}F{1}A{0}G{0}	HexNAc{2}Hex{3}Fuc{1}	H3N2F1
4	{N(N{H(H){H(H){H}}}}	H{4}N{2}	N{2}H{4}F{0}A{0}G{0}	HexNAc{2}Hex{4}	H4N2
5	{N(N{H(H){H(H{H}}}}}	H{4}N{2}	N{2}H{4}F{0}A{0}G{0}	HexNAc{2}Hex{4}	H4N2
6	{N(N{H(H){H}{N}}}}	H{3}N{3}	N{3}H{3}F{0}A{0}G{0}	HexNAc{3}Hex{3}	H3N3
7	{N(N{H(H){H(N}}}}	H{3}N{3}	N{3}H{3}F{0}A{0}G{0}	HexNAc{3}Hex{3}	H3N3
8	{N(F){N(H{H}{H(H)}}}}	H{4}N{2}F{1}	N{2}H{4}F{1}A{0}G{0}	HexNAc{2}Hex{4}Fuc{1}	H4N2F1
9	{N(N{H(H){H(H){H}}}}}	H{5}N{2}	N{2}H{5}F{0}A{0}G{0}	HexNAc{2}Hex{5}	H5N2
10	{N(N{H(H){H(H{H}}}}}	H{5}N{2}	N{2}H{5}F{0}A{0}G{0}	HexNAc{2}Hex{5}	H5N2
11	{N(N{H(H{H}){H{H}}}}}	H{5}N{2}	N{2}H{5}F{0}A{0}G{0}	HexNAc{2}Hex{5}	H5N2
12	{N(F){N{H{H}{H{N}}}}}	H{3}N{3}F{1}	N{3}H{3}F{1}A{0}G{0}	HexNAc{3}Hex{3}Fuc{1}	H3N3F1
13	{N(N{H{H}{H(N(F)}}}}}	H{3}N{3}F{1}	N{3}H{3}F{1}A{0}G{0}	HexNAc{3}Hex{3}Fuc{1}	H3N3F1
14	{N(F){N{H{H}{H}(N)}}}	H{3}N{3}F{1}	N{3}H{3}F{1}A{0}G{0}	HexNAc{3}Hex{3}Fuc{1}	H3N3F1
15	{N(N{H{H{H}}{H(N)}}}}	H{4}N{3}	N{3}H{4}F{0}A{0}G{0}	HexNAc{3}Hex{4}	H4N3
16	{N(N{H{H}{N}{H{H}}}}}	H{4}N{3}	N{3}H{4}F{0}A{0}G{0}	HexNAc{3}Hex{4}	H4N3
17	{N(N{H{H}{H(N{H}}}}}}	H{4}N{3}	N{3}H{4}F{0}A{0}G{0}	HexNAc{3}Hex{4}	H4N3
18	{N(N{H{H}{N}{H(N}}}}}	H{3}N{4}	N{4}H{3}F{0}A{0}G{0}	HexNAc{4}Hex{3}	H3N4
19	{N(N{H{H(N}}{H(N}}}}}	H{3}N{4}	N{4}H{3}F{0}A{0}G{0}	HexNAc{4}Hex{3}	H3N4

FIG. 10

55137834

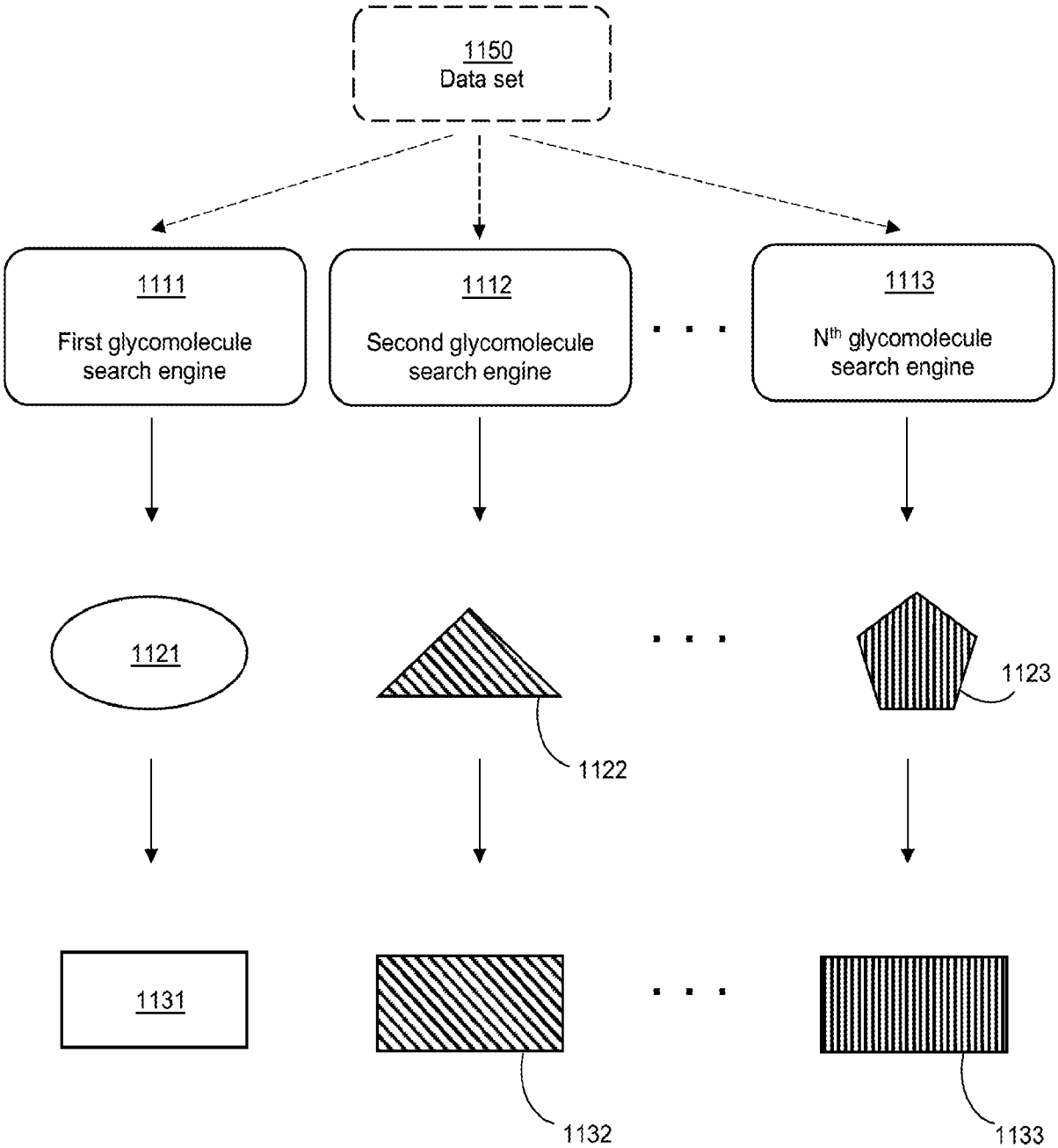

1150
Data set
1111
First glycomolecule search engine
1112
Second glycomolecule search engine
1113
Nth glycomolecule search engine
1121
1122
1123
1131
1132
1133

FIG. 11

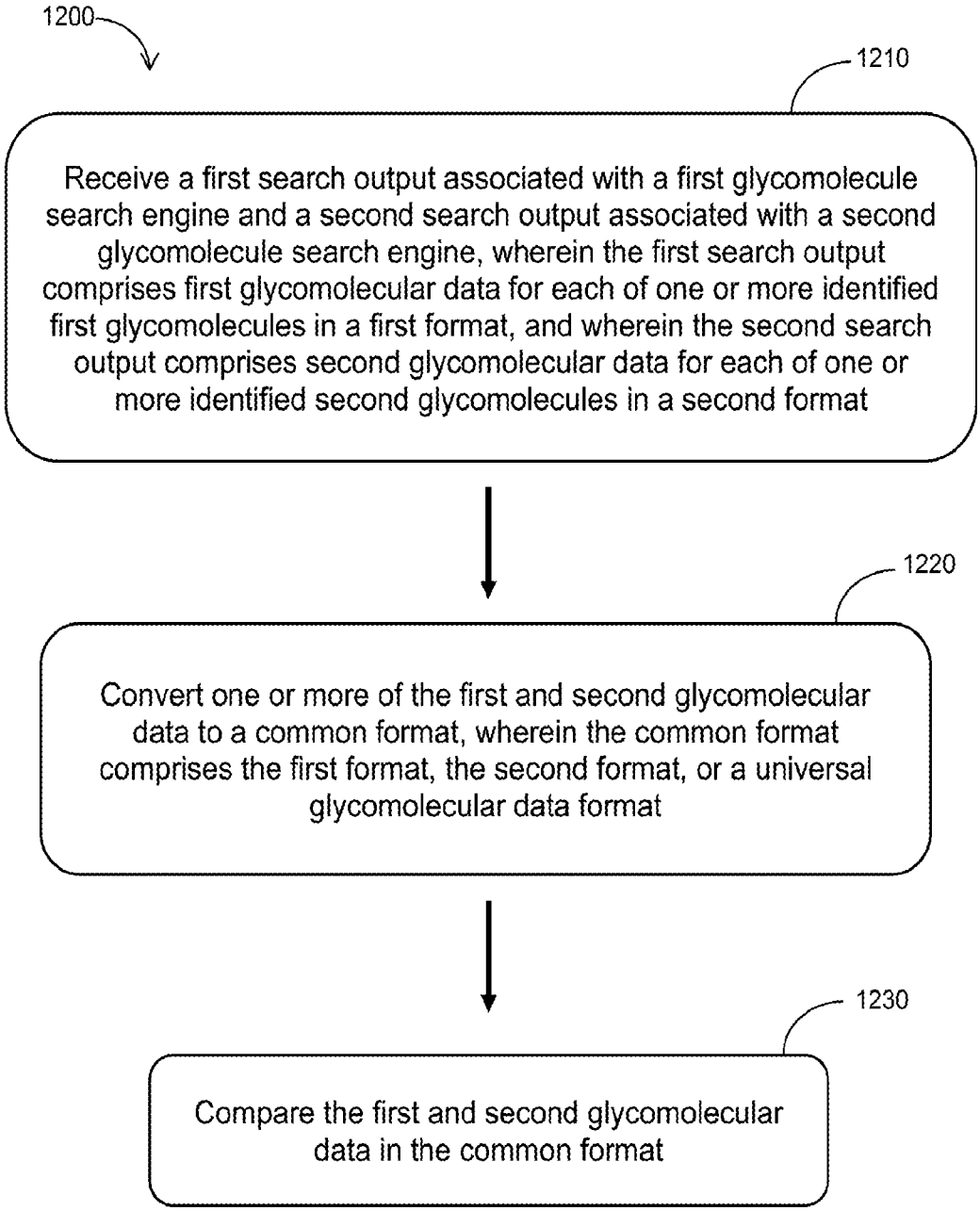

1200
1210
Receive a first search output associated with a first glycomolecule search engine and a second search output associated with a second glycomolecule search engine, wherein the first search output comprises first glycomolecular data for each of one or more identified first glycomolecules in a first format, and wherein the second search output comprises second glycomolecular data for each of one or more identified second glycomolecules in a second format
1220
Convert one or more of the first and second glycomolecular data to a common format, wherein the common format comprises the first format, the second format, or a universal glycomolecular data format
1230
Compare the first and second glycomolecular data in the common format

FIG. 12

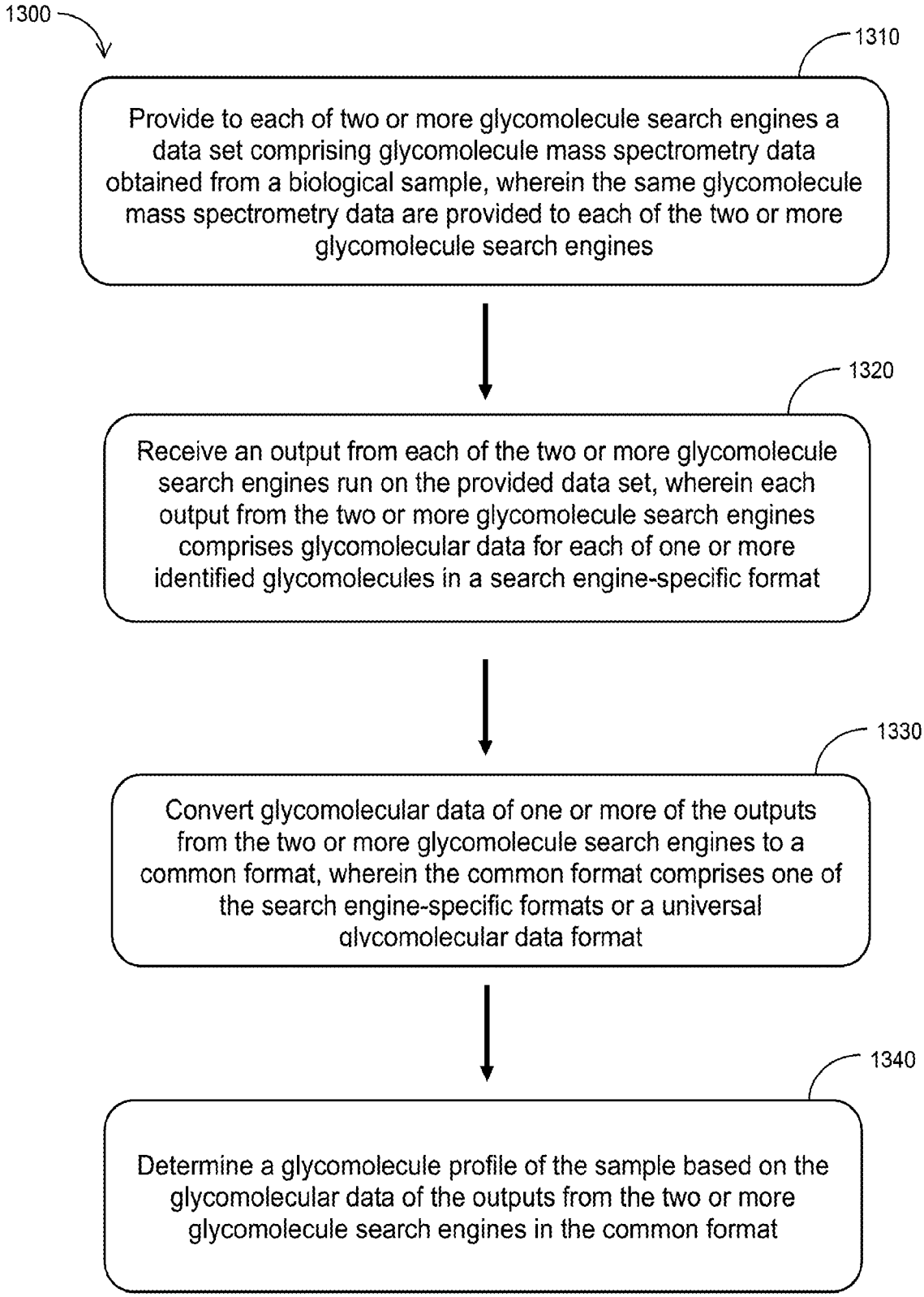

1300
1310
Provide to each of two or more glycomolecule search engines a data set comprising glycomolecule mass spectrometry data obtained from a biological sample, wherein the same glycomolecule mass spectrometry data are provided to each of the two or more glycomolecule search engines
1320
Receive an output from each of the two or more glycomolecule search engines run on the provided data set, wherein each output from the two or more glycomolecule search engines comprises glycomolecular data for each of one or more identified glycomolecules in a search engine-specific format
1330
Convert glycomolecular data of one or more of the outputs from the two or more glycomolecule search engines to a common format, wherein the common format comprises one of the search engine-specific formats or a universal glycomolecular data format
1340
Determine a glycomolecule profile of the sample based on the glycomolecular data of the outputs from the two or more glycomolecule search engines in the common format

FIG. 13

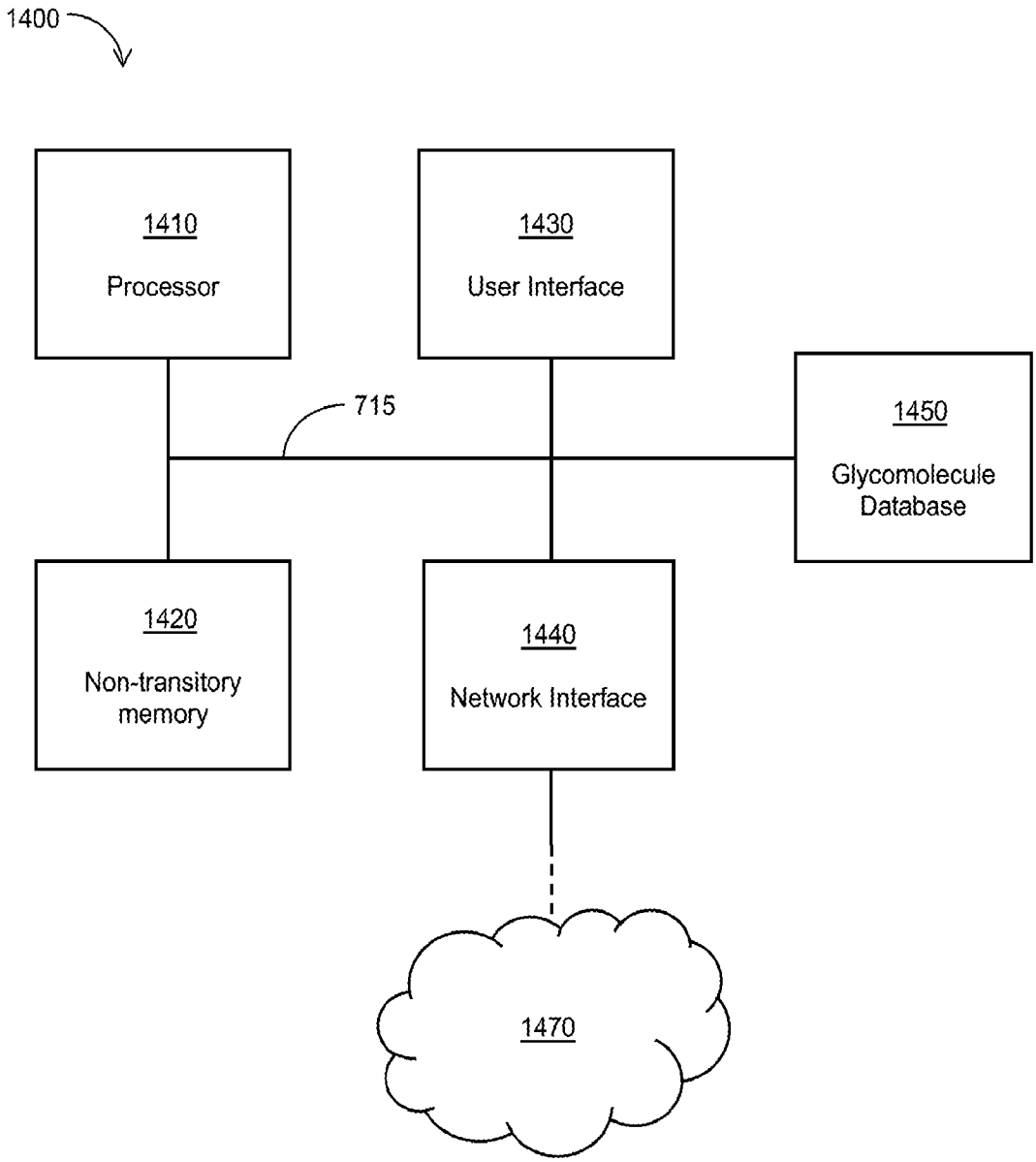

1400
1410
Processor
1430
User Interface
715
1450
Glycomolecule Database
1420
Non-transitory memory
1440
Network Interface
1470

FIG. 14

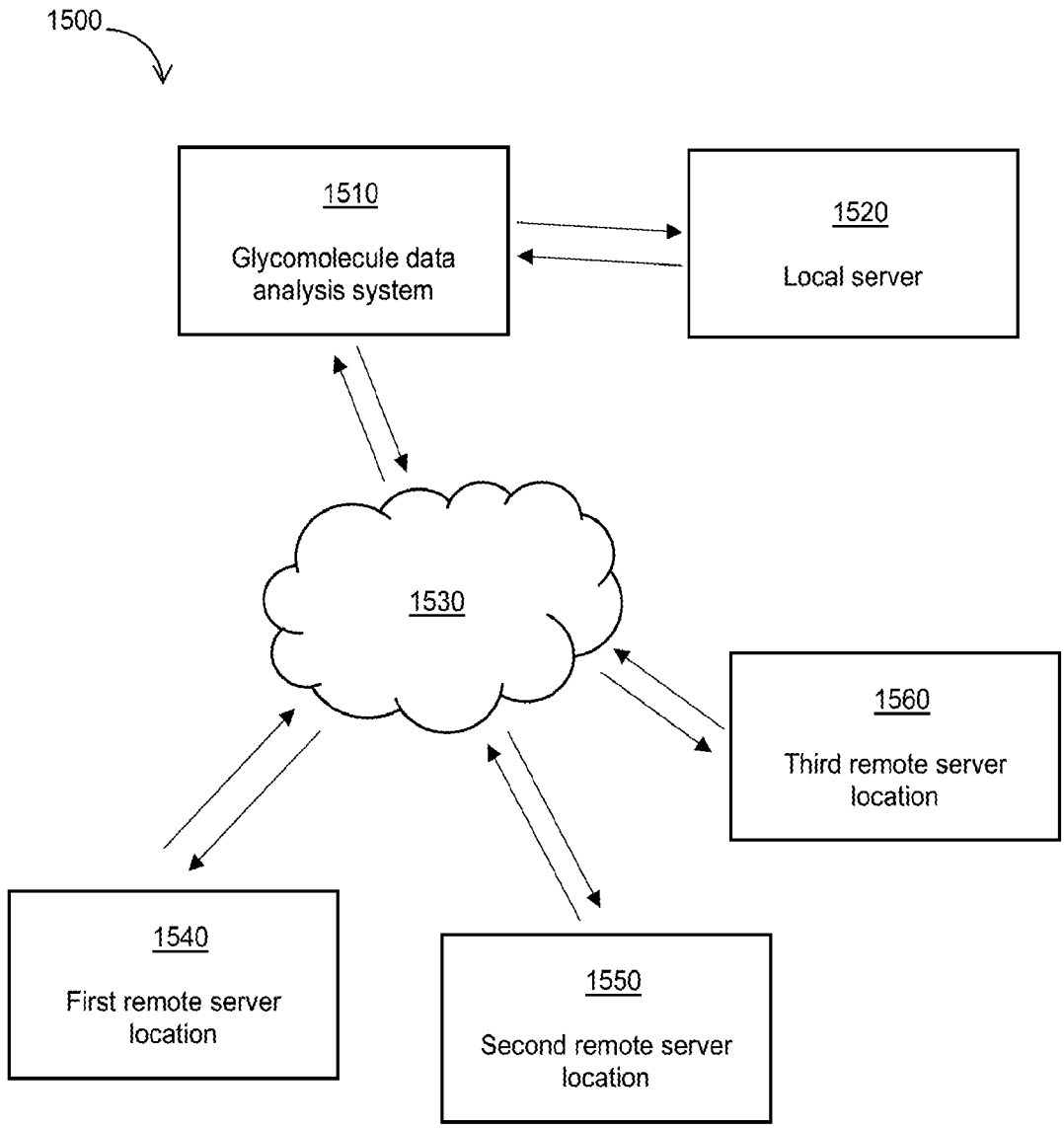
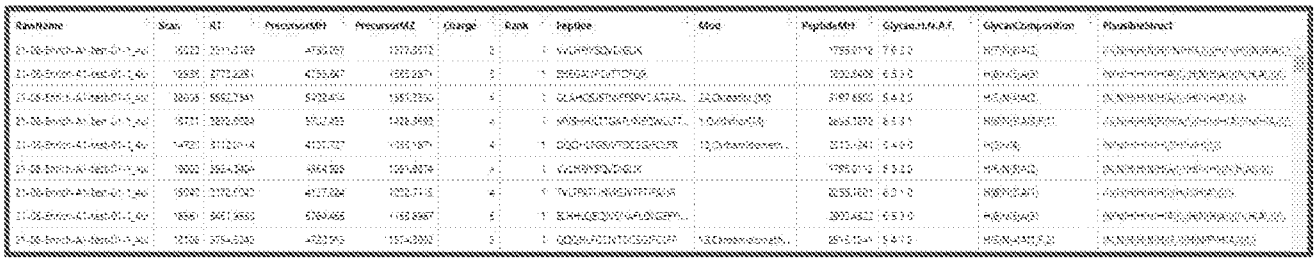

1500
1510
Glycomolecule data analysis system
1520
Local server
1530
1560
Third remote server location
1540
First remote server location
1550
Second remote server location

FIG. 15

pGlyco results



RawName	Scan	RT	PrecursorMH	PrecursorMZ	Charge	Rank	Peptide	Miss	PeptideMH	GlycanAtAA.P.	GlycanComposition	PlausibleStruct
21-06-Enrich-A1-ben-01-1_40	10223	2511.0169	4750.1057	1317.0013	3	1	VVLSPPSQVLVSLK		1395.8116	7.6 2.0	H(7)N(5)A(2)	
21-06-Enrich-A1-test-01-1_40	10535	2772.6281	4755.067	1585.0979	3	1	EVSDAVPVGVTCPQK		1650.8408	6.5 3.0	H(6)N(5)A(3)	
21-06-Enrich-A1-test-01-1_40	10656	2862.7581	5439.0476	1360.2556	4	1	QLAAQSQSTNTSSPVLATSPA...	2xOxidation (M)	2197.8800	5.4 2.5	H(5)N(4)A(2)	
21-06-Enrich-A1-ben-01-1_40	10711	2891.0583	5032.455	1408.8988	4	1	VNISHHGTTSMAPLRGTSNLGT...	1 Oxidation (M)	2668.3272	6.5 3.1	H(6)N(5)A(3)F(1)	
21-06-Enrich-A1-test-01-1_40	10750	2913.0514	4127.727	1032.9371	4	1	DQQHLSGGSVFIFPPKIKDVLM	1xCarbamidometh...	2215.1041	5.4 4.0	H(5)A(4)	
21-06-Enrich-A1-test-01-1_40	10802	2954.5404	4964.505	1655.8274	3	1	VVLSPPSQVLVSLK		1395.8116	5.5 2.5	H(5)N(5)A(2)	
21-06-Enrich-A1-ben-01-1_40	10940	3070.0742	4127.004	1032.7719	4	1	TVLFPPSTLNVSSVTTPSLVR		2255.1001	6.0 1.0	H(6)N(5)A(2)	
21-06-Enrich-A1-test-01-1_40	10956	3083.9668	5760.4455	1152.6987	5	1	SLKHHQESNVSFAFLDNSEPP...		3200.4622	6.5 3.0	H(6)N(5)A(3)	
21-06-Enrich-A1-ben-01-1_40	10986	3104.6248	4083.995	1361.6722	3	1	QQQHLFGSNVTDCSGFPSSAT	1xCarbamidometh...	2814.1264	5.4 1.0	H(5)N(4)A(3)F(1)	

Byonic results

	Query #2	Protein Rank	Pep Sequence	Glycoform	Modification type(s)	Observed m/z	Z	Observed (M+H)	mass (M+H)	lgmass error	error (ppm)	Starting position	Cleavage	Score
1	338880	1	K.VPPLSLSLSQTVSDQSNVVVACL+57.021[C]VQGPPKQEPMDSP.N[+2366...	HexNAc(2)Hex(5)HexNAc(1)Hex3Fuc(1)	C[+57], N[+2366]	1586.0273	6	9575.920	9575.919	0	2.6	18	Specific	9.55
2	370390	1	K.VPPLSLSLSQTVSDQSNVVVACL+57.021[C]VQGPPKQEPMDSP.N[+2075...	HexNAc(4)Hex(5)Fuc(1)NeuGc(1)	C[+57], N[+2075]	1911.7599	6	9560.790	9560.780	0	0.6	18	Specific	3.61
3	450776	1	K.VPPLSLSLSQTVSDQSNVVVACL+57.021[C]VQGPPKQEPMDSP.N[+2366...	HexNAc(2)Hex(5)Fuc(1)NeuGc(2)	C[+57], N[+2366]	1592.5655	6	9572.845	9572.859	0	2.5	18	Specific	22.65
4	396886	1	K.VPPLSLSLSQTVSDQSNVVVACL+57.021[C]VQGPPKQEPMDSP.N[+1747...	HexNAc(3)Hex(5)Fuc(1)NeuGc(2)	C[+57], N[+1748]	1118.8286	6	6707.980	6707.983	0	0.3	18	Specific	49.79
5	450226	1	K.VPPLSLSLSQTVSDQSNVVVACL+57.021[C]VQGPPKQEPMDSP.N[+2366...	HexNAc(4)Hex(5)Fuc(1)NeuGc(2)	C[+57], N[+2366]	1982.5111	6	9571.855	9572.851	-1	0.4	18	Specific	50.22
6	417560	1	K.VPPLSLSLSQTVSDQSNVVVACL+57.021[C]VQGPPKQEPMDSP.N[+2366...	HexNAc(2)Hex(5)Fuc(2)NeuGc(2)	C[+57], N[+2366]	1662.9784	6	9972.840	9972.851	0	1.5	18	Specific	75.77
7	411076	1	K.VPPLSLSLSQTVSDQSNVVVACL+57.021[C]VQGPPKQEPMDSP.N[+2366...	HexNAc(4)Hex(5)Fuc(1)NeuGc(3)	C[+57], N[+2366]	1662.6139	6	9971.841	9972.851	-1	2.1	18	Specific	65.42
8	408316	1	K.VPPLSLSLSQTVSDQSNVVVACL+57.021[C]VQGPPKQEPMDSP.N[+2366...	HexNAc(2)Hex(5)Fuc(2)NeuGc(3)	C[+57], N[+2366]	1663.1456	6	9973.856	9972.851	1	0.1	18	Specific	88.66
9	399368	1	K.VPPLSLSLSQTVSDQSNVVVACL+57.021[C]VQGPPKQEPMDSP.N[+1566...	HexNAc(3)Hex(6)NeuAc(1)NeuGc(1)	C[+57], N[+1566]	1058.2259	6	6375.319	6375.319	0	1.3	18	Specific	102.08
10	365226	1	K.VPPLSLSLSQTVSDQSNVVVACL+57.021[C]VQGPPKQEPMDSP.N[+2075...	HexNAc(4)Hex(5)Fuc(1)NeuGc(1)	C[+57], N[+2075]	1911.7573	6	9560.746	9560.780	0	3.6	18	Specific	121.55
11	408890	1	K.VPPLSLSLSQTVSDQSNVVVACL+57.021[C]VQGPPKQEPMDSP.N[+2366...	HexNAc(2)Hex(5)Fuc(2)NeuGc(2)	C[+57], N[+2366]	1662.9784	6	9972.836	9972.851	0	0.5	18	Specific	137.45

FIG. 16

pGlyco3 results

	GlySpec	PepSpec	RawName	Scan	RT	PrecursorMH	PrecursorMZ	Charge
2	SS_HILIC_20	SS_HILIC_20	SS_HILIC_20	48430	4746.74095	4569.95665	1143.24462	4
3	SS_HILIC_20	SS_HILIC_20	SS_HILIC_20	31255	3183.08482	5106.05254	1022.01633	5
4	SS_HILIC_20	SS_HILIC_20	SS_HILIC_20	36153	3636.55583	4719.88087	1180.72567	4
5	SS_HILIC_20	SS_HILIC_20	SS_HILIC_20	36944	3709.9834	5074.06822	1269.27251	4
6	SS_HILIC_20	SS_HILIC_20	SS_HILIC_20	28387	2919.66971	4723.01779	945.40938	5
7	SS_HILIC_20	SS_HILIC_20	SS_HILIC_20	17428	1913.44388	3152.21264	1051.40907	3

pGlyco3 result to Universal glycopeptide format

	MS2 #	Protein ID	Peptide Sequen	Glycan	Precursor MZ	Z	RT	Total Score	Search Engine
2	48430	sp|P02765|FET	AALAAFNAQNN	N(4)H(5)F(0)A(2)G(0)	1143.24462	4	4746.74095	6.75064	pGlyco3
3	31255	sp|P01011|AA(	TLNQSSDELQLS	N(5)H(6)F(0)A(3)G(0)	1022.01633	5	3183.08482	9.30088	pGlyco3
4	36153	sp|P02787|TRF	QQQHLFGSNVT	N(4)H(5)F(0)A(2)G(0)	1180.72567	4	3636.55583	19.0117	pGlyco3
5	36944	sp|P01011|AA(	TLNQSSDELQLS	N(5)H(6)F(0)A(3)G(0)	1269.27251	4	3709.9834	57.09853	pGlyco3
6	28387	sp|Q14624|ITIH	LPTQNITFQTES	N(4)H(5)F(0)A(1)G(0)	945.40938	5	2919.66971	34.75036	pGlyco3
7	17428	sp|P10909|CLU	EDALNETR	N(4)H(5)F(0)A(2)G(0)	1051.40907	3	1913.44388	45.93505	pGlyco3
8	12800	sp|P0DOX5|IG(	EEQYNSTYR	N(4)H(5)F(0)A(1)G(0)	1035.06885	3	1495.21678	18.84852	pGlyco3

FIG. 17

Byonic results

	A	B	C	D	E	F	G	H	I	K	L	M	N
1	Query #s	Protein Rank	Peptide <ProteinMetrics Confidential>	Glycans NHFAGNa	Modification Type(s)	Observed m/z	z	Observed (M+H)	Calc. mass (M+H)	Off-by-x error	Mass error (ppm)	Starting position	Cleavage	Score
2	12379:3	1	R.NEEYN[+3861.380]K.S	HexNAc(5)Hex(6)Fu	N[+3861]	1219.788	3	3657.349	3657.347	0	0.5	52	Specific	38.8
3	10288:3	1	R.NEEYN[+2407.852]K.S	HexNAc(5)Hex(5)Fu	N[+2403]	1068.735	3	3204.192	3204.185	0	2.3	52	Specific	51.1
4	07334:3	1	R.NEEYN[+2569.905]K.S	HexNAc(5)Hex(6)Fu	N[+2579]	1122.758	3	3366.259	3366.252	0	2.1	52	Specific	75.5
5	07080:2	1	R.NEEYN[+1913.677]K.S	HexNAc(4)Hex(5)Fu	N[+1914]	1355.526	2	2710.032	2710.024	0	2.8	52	Specific	83.8
6	09077:3	1	R.NEEYN[+2715.963]K.S	HexNAc(5)Hex(6)Fu	N[+2716]	1171.447	3	3513.326	3513.319	0	1.8	52	Specific	84.5
7	07145:2	1	R.NEEYN[+2204.771]K.S	HexNAc(4)Hex(5)Fu	N[+2205]	1001.047	3	3003.127	3003.120	0	2.5	52	Specific	87.3

Byonic result to Universal glycopeptide format

	A	B	C	D	E	F	G	H	I
1	MS2 #	Protein ID	Peptide Sequence	Glycan	Precursor MZ	Z	RT (Minutes)	Total Score	Search Engine
2	14691	P02763	NEEYNK	N(5)H(6)F(0)A(3)G(0)	1219.7879	3	27.7244	38.84	Byonic
3	12481	P02763	NEEYNK	N(5)H(5)F(0)A(2)G(0)	1068.7354	3	24.4681	51.12	Byonic
4	9337	P02763	NEEYNK	N(5)H(6)F(0)A(2)G(0)	1122.7578	3	20.1784	75.45	Byonic
5	9062	P02763	NEEYNK	N(4)H(5)F(0)A(1)G(0)	1355.5196	2	19.8369	83.77	Byonic
6	12205	P02763	NEEYNK	N(5)H(6)F(1)A(2)G(0)	1171.4468	3	24.0657	84.45	Byonic
7	9132	P02763	NEEYNK	N(4)H(5)F(0)A(2)G(0)	1001.0473	3	19.9254	87.31	Byonic
8	12755	P02763	NEEYNK	N(4)H(5)F(0)A(2)G(0)	1501.0672	2	24.8586	89.03	Byonic

FIG. 18

Metamorpheus results

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	File Name	Scan Nu	Scan Rel	Precursi	Precursi	Precursi	Precursi	Protein	Organist	Protein	Start an	Base Sec	Flanking	Full Seq	Numbe	Peptide
2	D:\ChiHung\	18233	35.78305	18200	1093.71	4	4370.811	P08603	Homo sapi	Compleme	[868 to 885	IPCSQPPQI	K,S	IPC[Comme	2	4370.81
3	D:\ChiHung\	15991	35.43841	15959	1093.711	4	4370.817	P08603	Homo sapi	Compleme	[868 to 885	IPCSQPPQI	K,S	IPC[Comme	2	4370.81
4	D:\ChiHung\	28836	53.28694	28804	944.0035	5	4714.981	P02765	Homo sapi	Alpha-2-HS	[166 to 182	AALAAFNAC	K,A	AALAAFNAC	1	4714.976
5	D:\ChiHung\	17339	37.87318	17306	1415.231	3	4242.671	P00450	Homo sapi	Ceruloplas	[129 to 144	EHEGAIYPD	K,A	EHEGAIYPD	1	4242.664
6	D:\ChiHung\	19280	40.8451	19245	1225.731	4	4898.896	P00450	Homo sapi	Ceruloplas	[129 to 144	EHEGAIYPD	K,A	EHEGAIYPD	1	4898.892
7	D:\ChiHung\	18802	37.45858	18763	1061.673	4	4242.661	P00450	Homo sapi	Ceruloplas	[129 to 144	EHEGAIYPD	K,A	EHEGAIYPD	1	4242.664

Metamorpheus result to Universal glycopeptide format

	A	B	C	D	E	F	G	H	I
1	MS2 #	Protein ID	Peptide Sequence	Glycan	Precursor MZ	Z	RT (Minutes)	Total Score	Search Engine
2	18233	P08603	IPCSQPPQIEHGTINSSR	N(4)H(5)F(1)A(2)G(0)	4370.8108	4	35.78305336	95.38236674	Metamorpheus
3	15991	P08603	IPCSQPPQIEHGTINSSR	N(4)H(5)F(1)A(2)G(0)	4370.8166	4	35.43840609	87.3415359	Metamorpheus
4	28836	P02765	AALAAFNAQNNGSNFQLEEISR	N(4)H(5)F(1)A(2)G(0)	4714.9811	5	53.28694106	84.47914702	Metamorpheus
5	17339	P00450	EHEGAIYPDNTTDFQR	N(4)H(5)F(1)A(2)G(0)	4242.6709	3	37.87317728	83.36870837	Metamorpheus
6	19280	P00450	EHEGAIYPDNTTDFQR	N(5)H(6)F(1)A(3)G(0)	4898.8956	4	40.84509764	82.40018427	Metamorpheus
7	18802	P00450	EHEGAIYPDNTTDFQR	N(4)H(5)F(1)A(2)G(0)	4242.6613	4	37.45857611	82.38982312	Metamorpheus
8	19729	P00450	EHEGAIYPDNTTDFQR	N(4)H(5)F(1)A(2)G(0)	4242.6704	4	38.19379393	81.42315027	Metamorpheus

FIG. 19

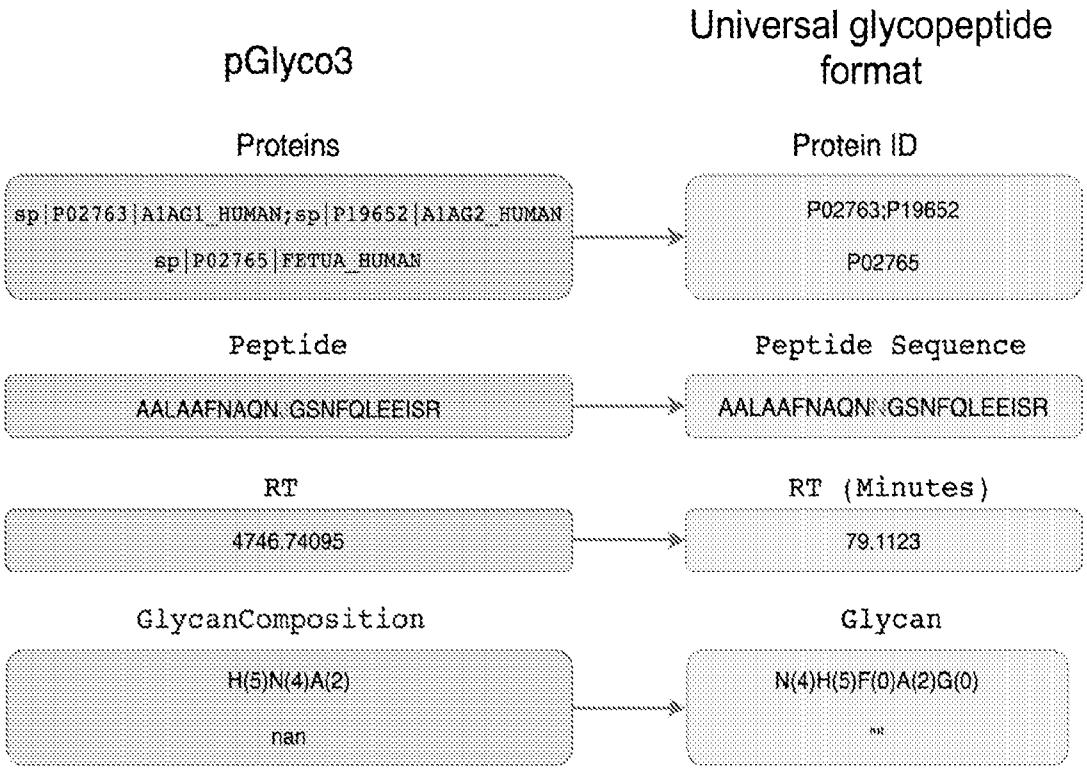

pGlyco3
Universal glycopeptide format
Proteins
sp|P02763|A1AG1_HUMAN;sp|P19652|A1AG2_HUMAN
sp|P02765|FETUA_HUMAN
Protein ID
P02763;P19652
P02765
Peptide
AALAAFNAQN.GSNFQLEEISR
Peptide Sequence
AALAAFNAQN.GSNFQLEEISR
RT
4746.74095
RT (Minutes)
79.1123
GlycanComposition
H(5)N(4)A(2)
nan
Glycan
N(4)H(5)F(0)A(2)G(0)

FIG. 20

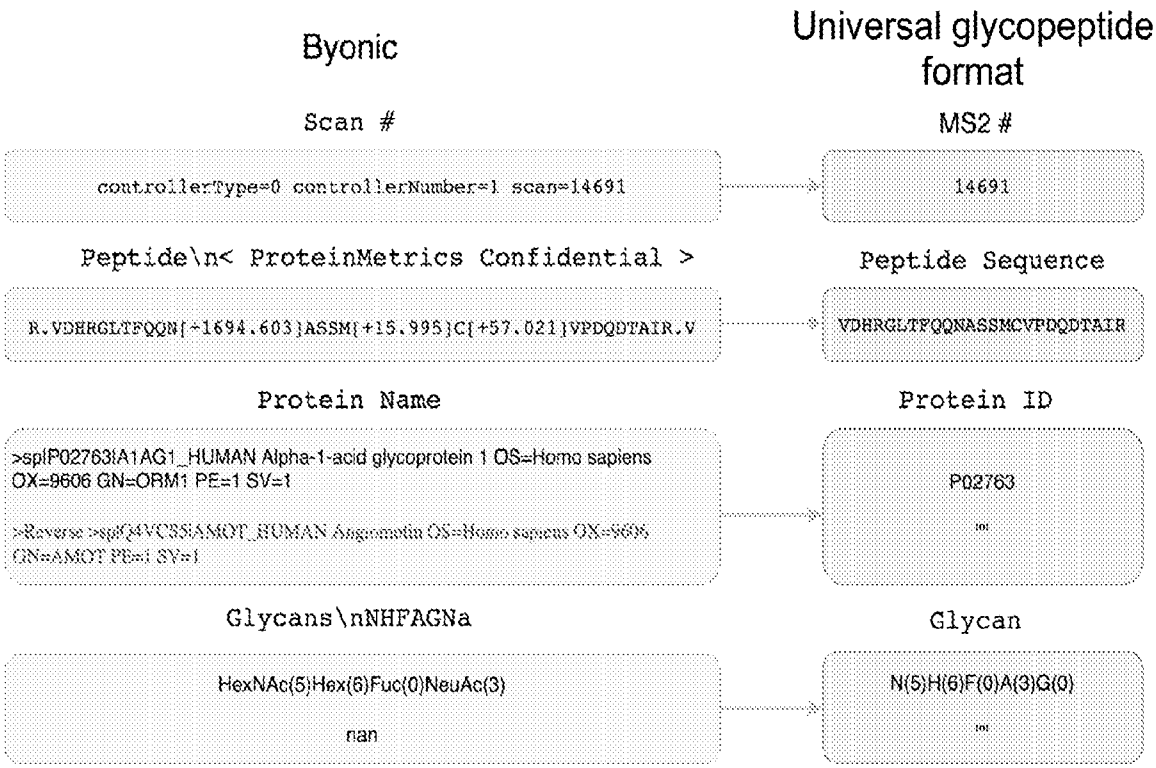

Byonic
Universal glycopeptide format
Scan #
MS2 #
controllerType=0 controllerNumber=1 scan=14691
14691
Peptide\n< ProteinMetrics Confidential >
Peptide Sequence
R.VDHRGLTFQQN[+1694.603]ASSM[+15.995]C[+57.021]VPDQDTAIR.V
VDHRGLTFQQNASSMCVPDQDTAIR
Protein Name
Protein ID
>sp|P02763|A1AG1_HUMAN Alpha-1-acid glycoprotein 1 OS=Homo sapiens OX=9606 GN=ORM1 PE=1 SV=1
>Reverse >sp|Q4VC33|AMOT_HUMAN Angiomotin OS=Homo sapiens OX=9606 GN=AMOT PE=1 SV=1
P02763
nan
Glycans\nNHFAGNa
Glycan
HexNAc(5)Hex(6)Fuc(0)NeuAc(3)
nan
N(5)H(6)F(0)A(3)G(0)
nan

FIG. 21

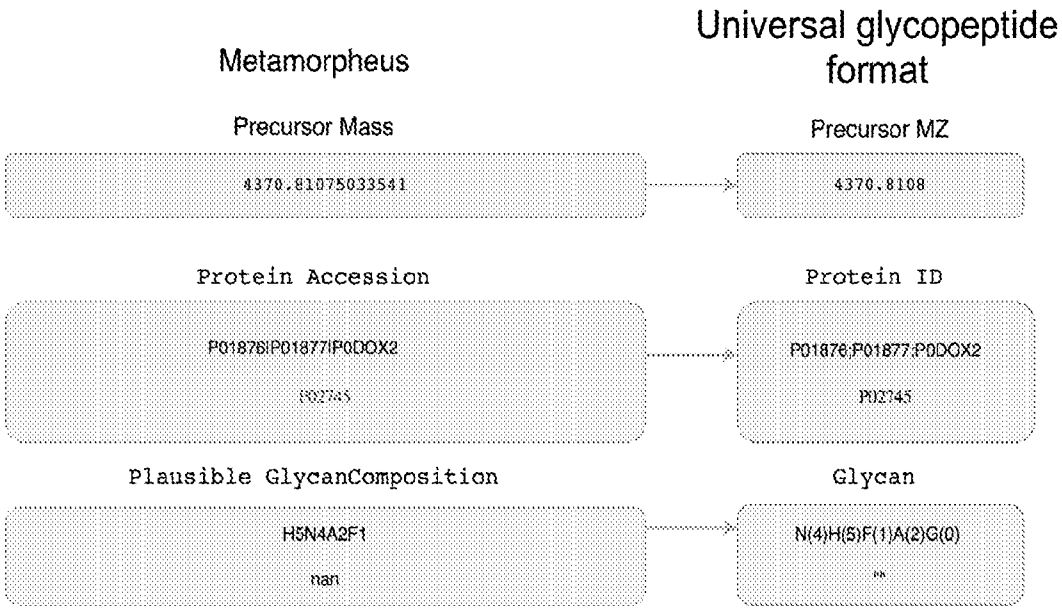

Metamorpheus
Universal glycopeptide format
Precursor Mass
4370.81075033541
Precursor MZ
4370.8108
Protein Accession
P01876|P01877|P0DOX2
P02745
Protein ID
P01876;P01877;P0DOX2
P02745
Plausible GlycanComposition
H5N4A2F1
nan
Glycan
N(4)H(5)F(1)A(2)G(0)

FIG. 22

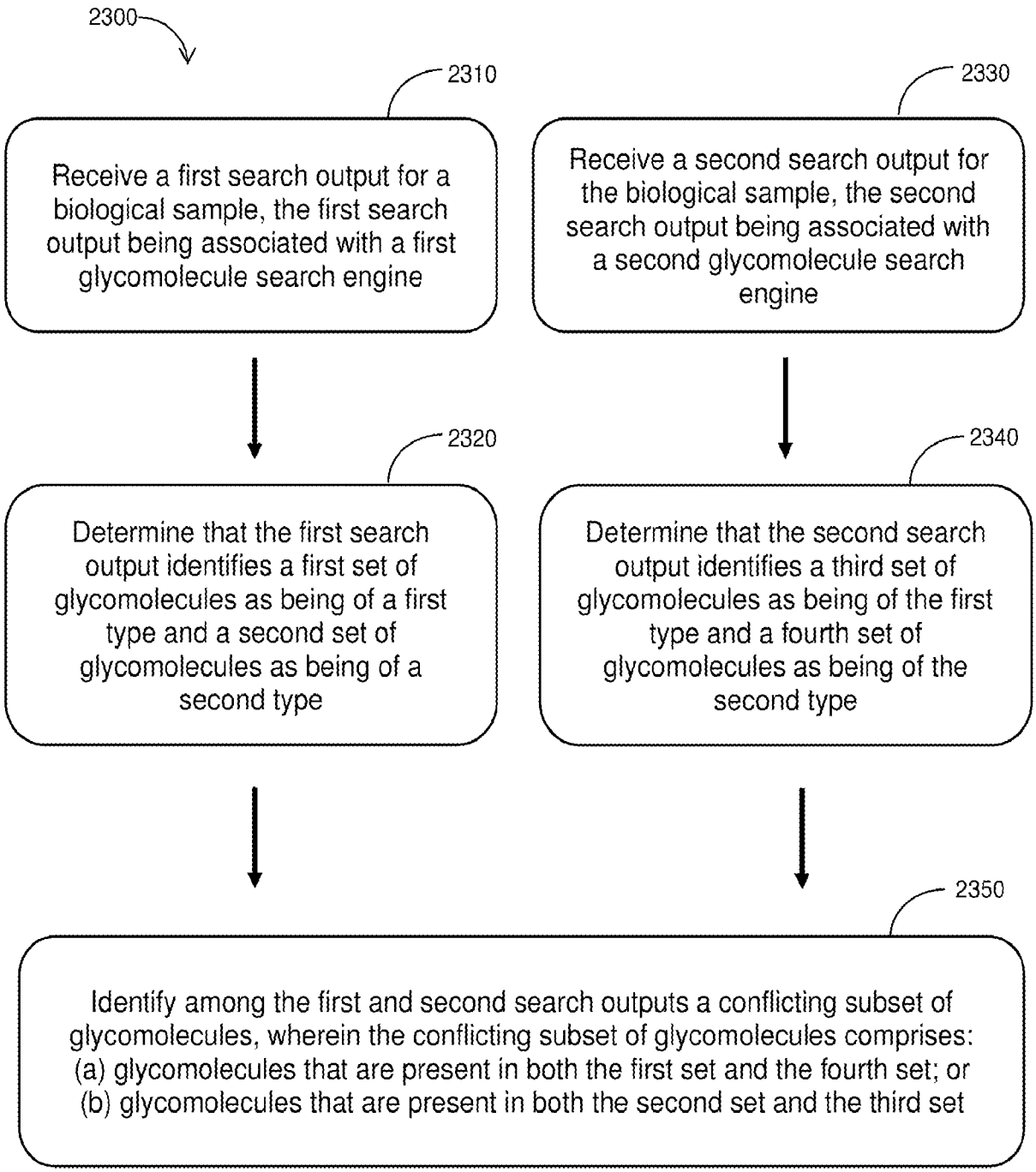

2300
2310
Receive a first search output for a biological sample, the first search output being associated with a first glycomolecule search engine
2330
Receive a second search output for the biological sample, the second search output being associated with a second glycomolecule search engine
2320
Determine that the first search output identifies a first set of glycomolecules as being of a first type and a second set of glycomolecules as being of a second type
2340
Determine that the second search output identifies a third set of glycomolecules as being of the first type and a fourth set of glycomolecules as being of the second type
2350
Identify among the first and second search outputs a conflicting subset of glycomolecules, wherein the conflicting subset of glycomolecules comprises: (a) glycomolecules that are present in both the first set and the fourth set; or (b) glycomolecules that are present in both the second set and the third set

FIG. 23

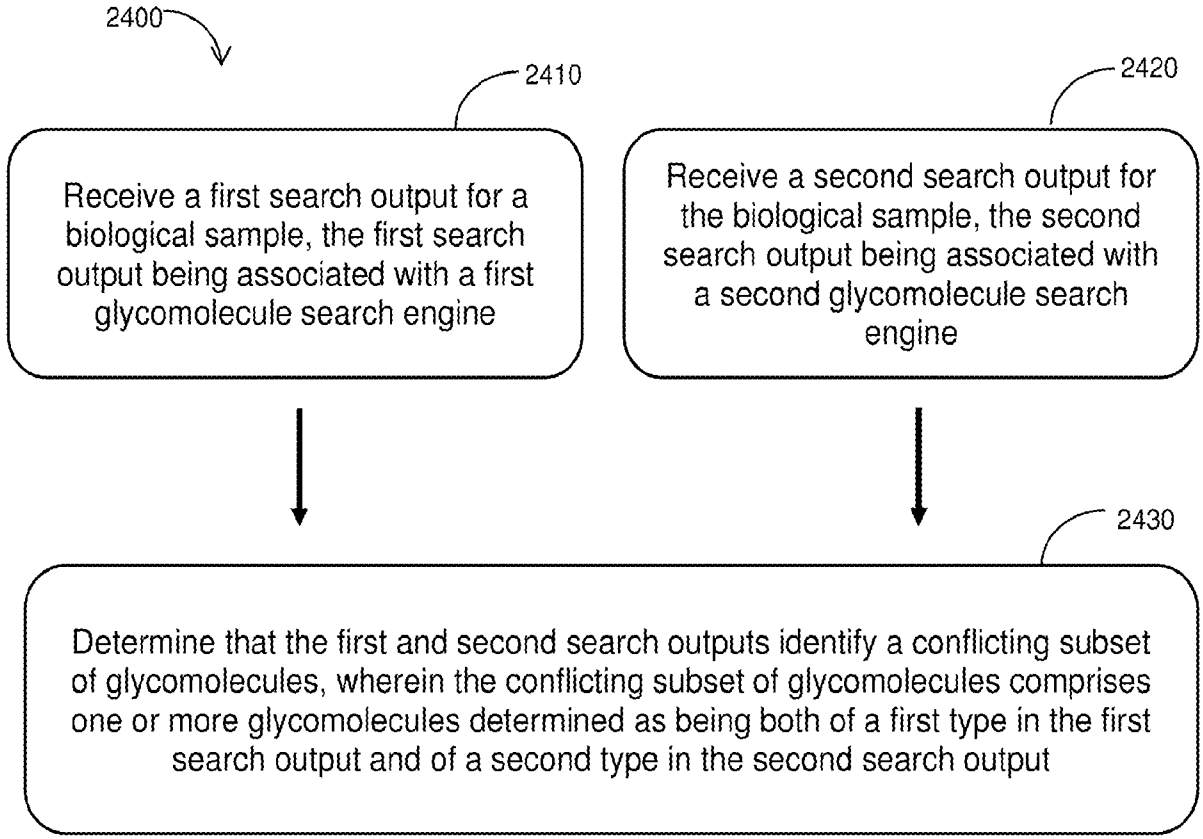

2400
2410
Receive a first search output for a biological sample, the first search output being associated with a first glycomolecule search engine
2420
Receive a second search output for the biological sample, the second search output being associated with a second glycomolecule search engine
2430
Determine that the first and second search outputs identify a conflicting subset of glycomolecules, wherein the conflicting subset of glycomolecules comprises one or more glycomolecules determined as being both of a first type in the first search output and of a second type in the second search output

FIG. 24A

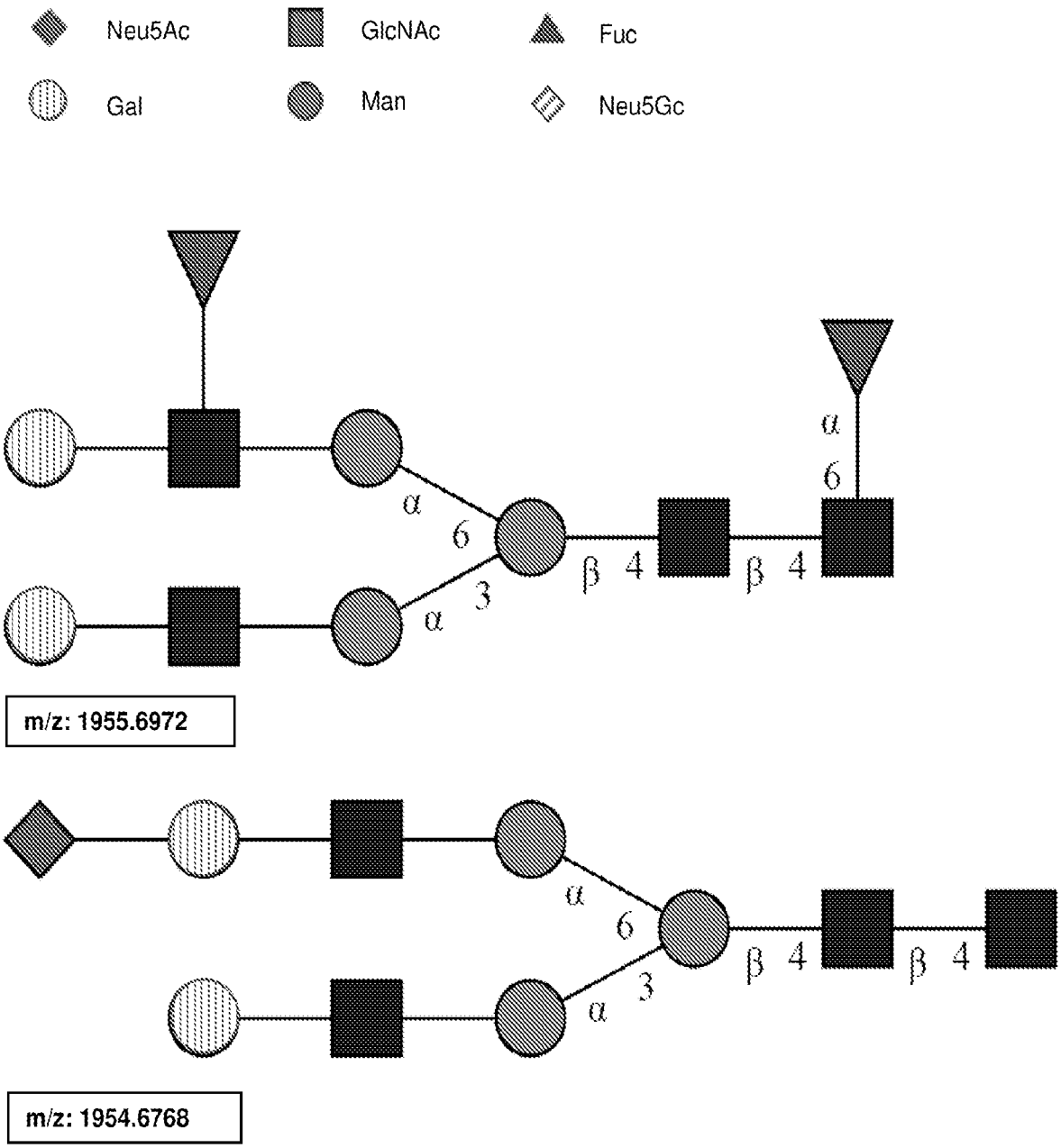


FIG. 24B

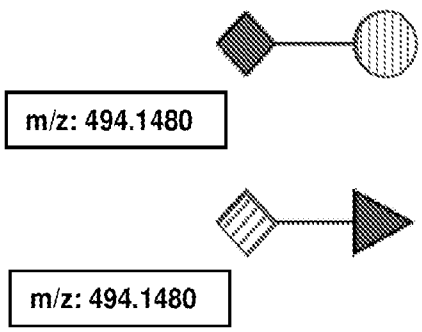


FIG. 24C

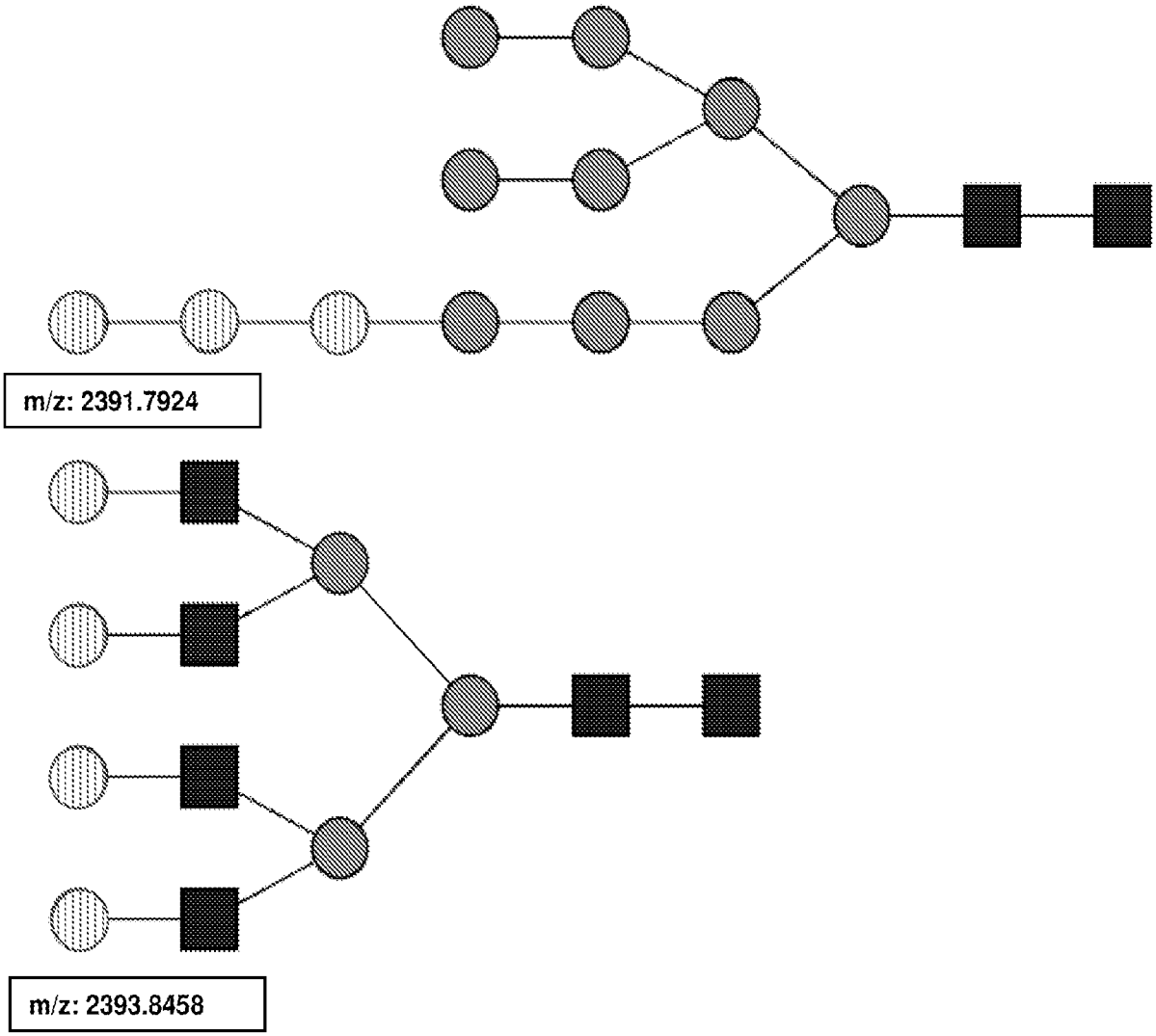


FIG. 24D

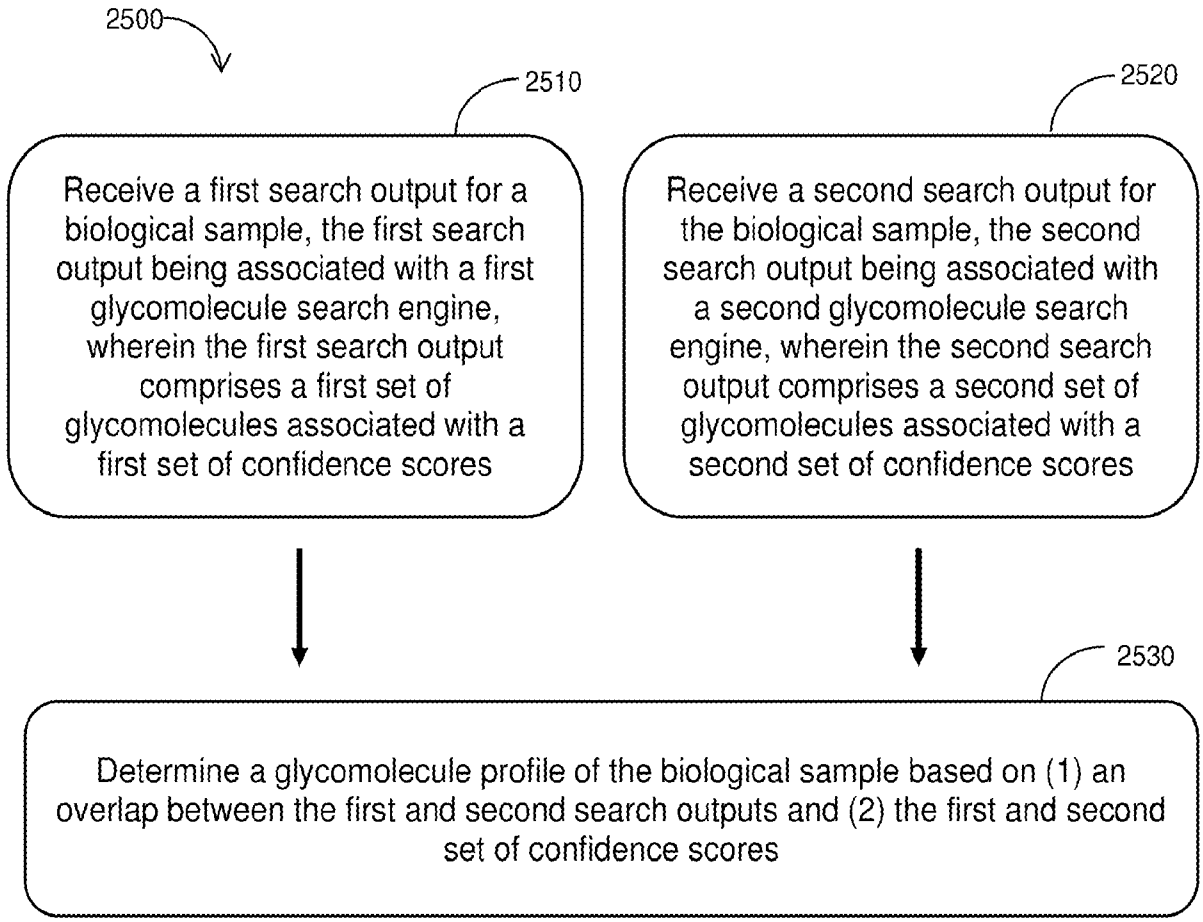

2500
2510
Receive a first search output for a biological sample, the first search output being associated with a first glycomolecule search engine, wherein the first search output comprises a first set of glycomolecules associated with a first set of confidence scores
2520
Receive a second search output for the biological sample, the second search output being associated with a second glycomolecule search engine, wherein the second search output comprises a second set of glycomolecules associated with a second set of confidence scores
2530
Determine a glycomolecule profile of the biological sample based on (1) an overlap between the first and second search outputs and (2) the first and second set of confidence scores

FIG. 25

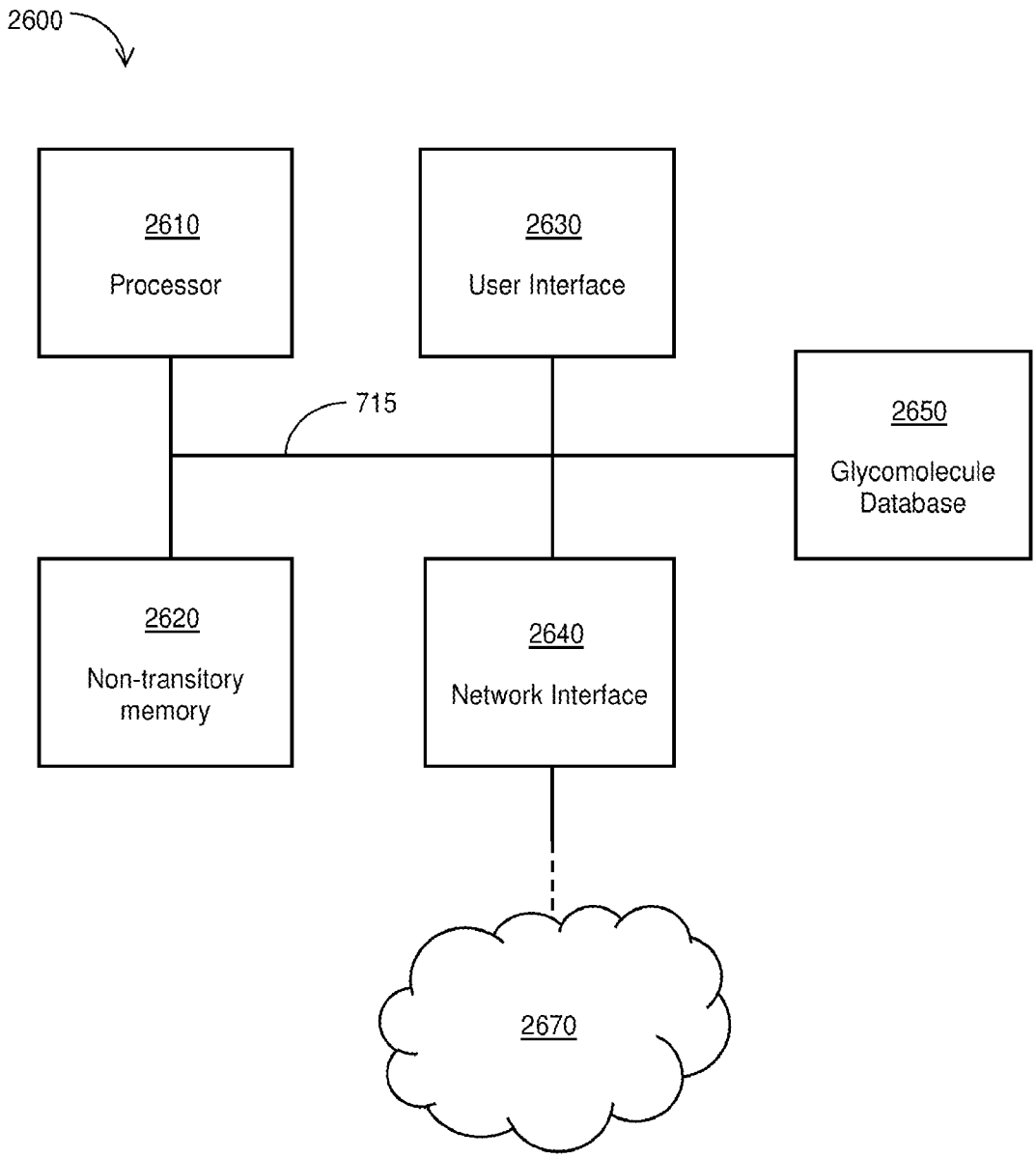

2600
2610
Processor
2630
User Interface
715
2650
Glycomolecule Database
2620
Non-transitory memory
2640
Network Interface
2670

FIG. 26

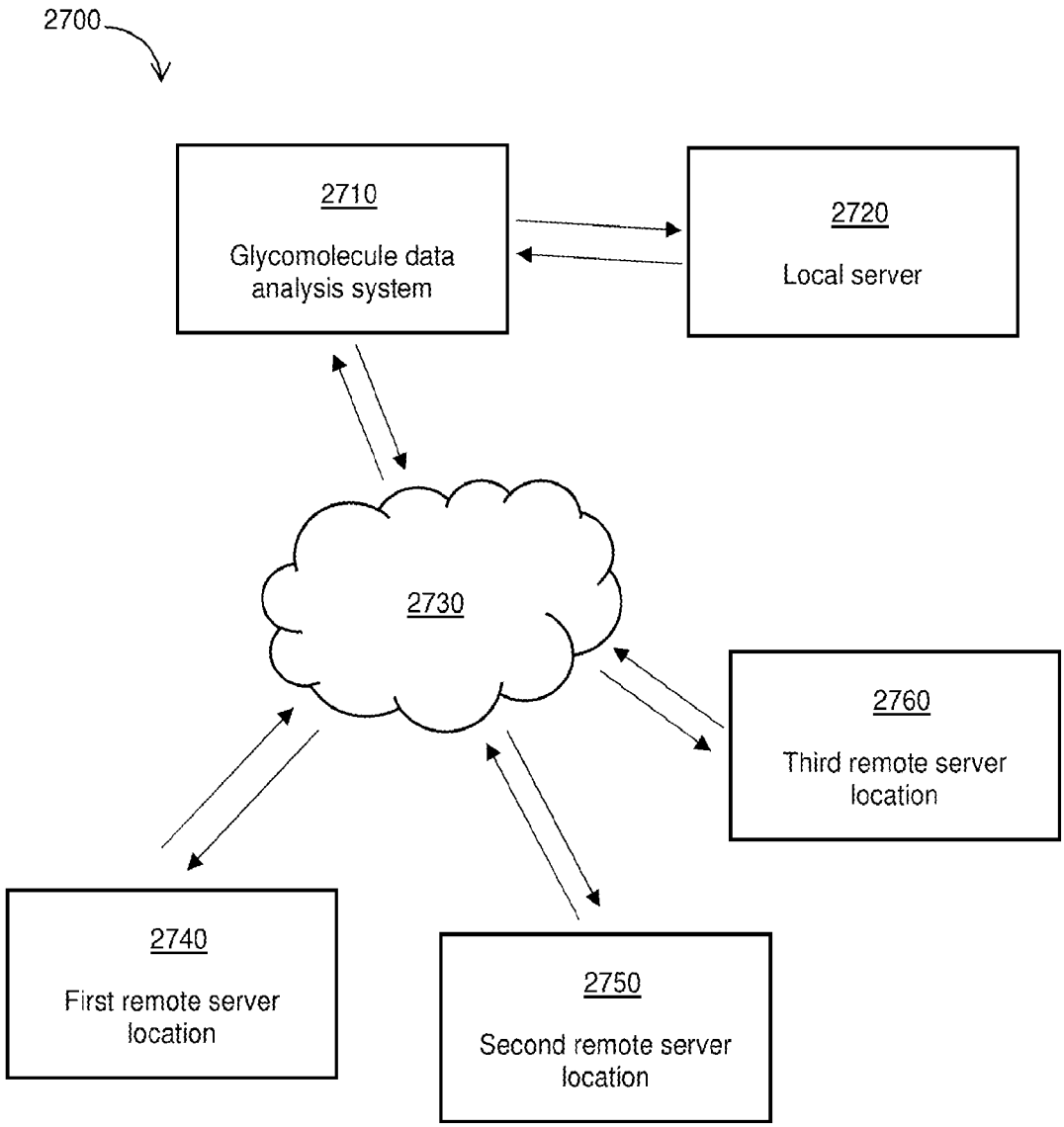

2700
2710
Glycomolecule data analysis system
2720
Local server
2730
2760
Third remote server location
2740
First remote server location
2750
Second remote server location

FIG. 27

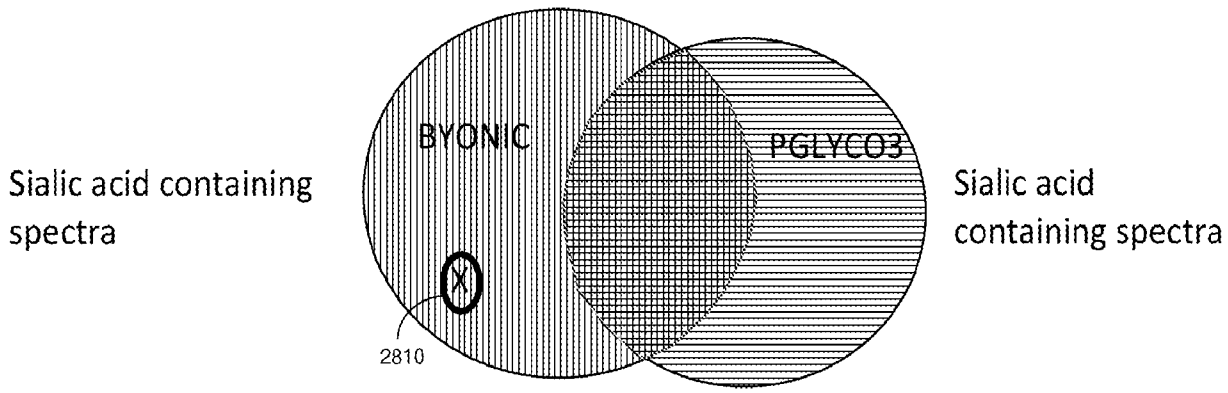


FIG. 28A

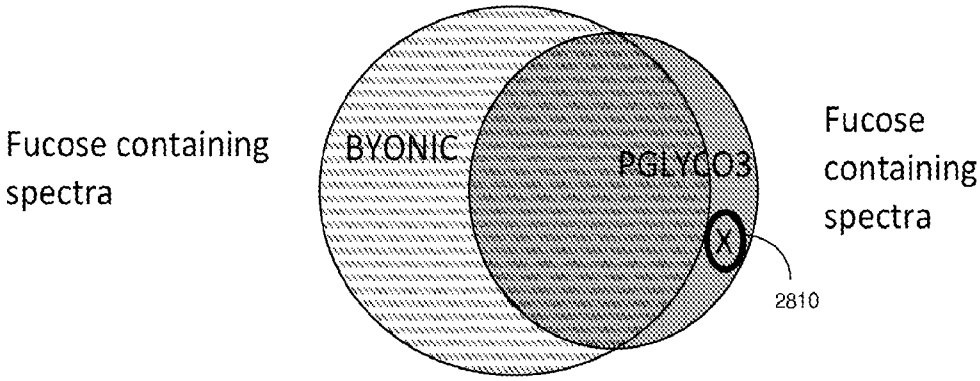


FIG. 28B

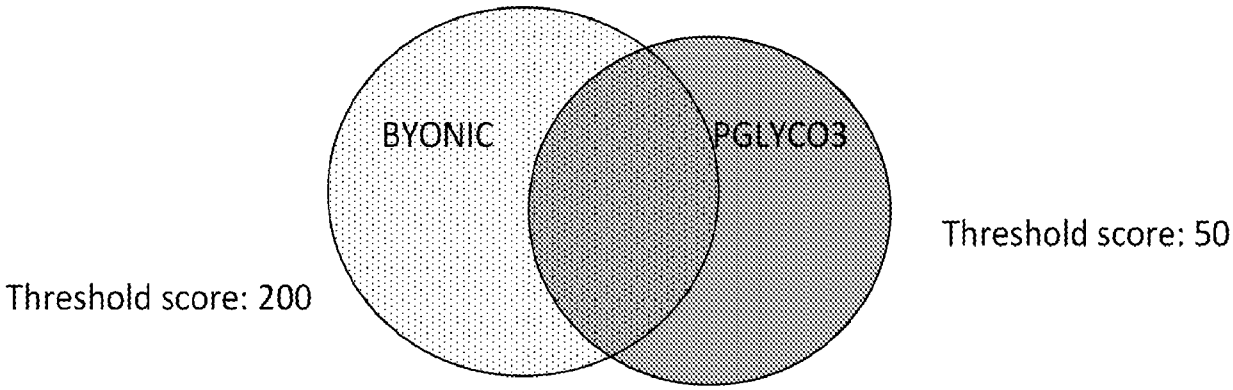


FIG. 29A

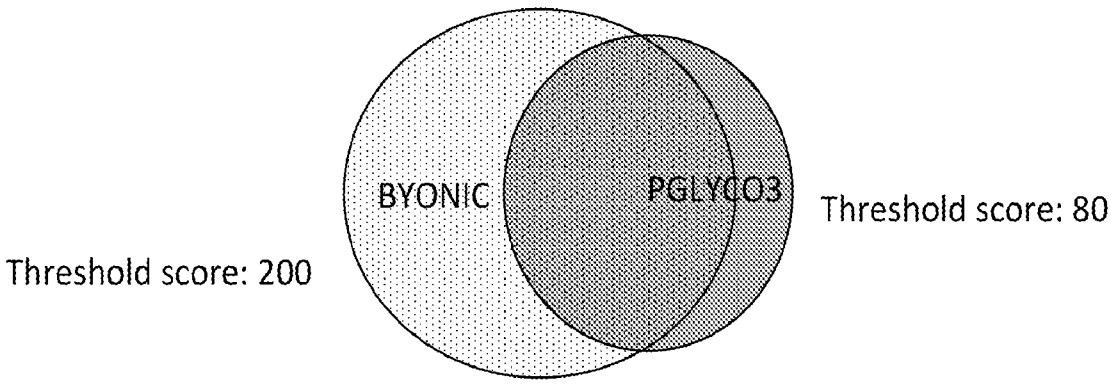


FIG. 29B

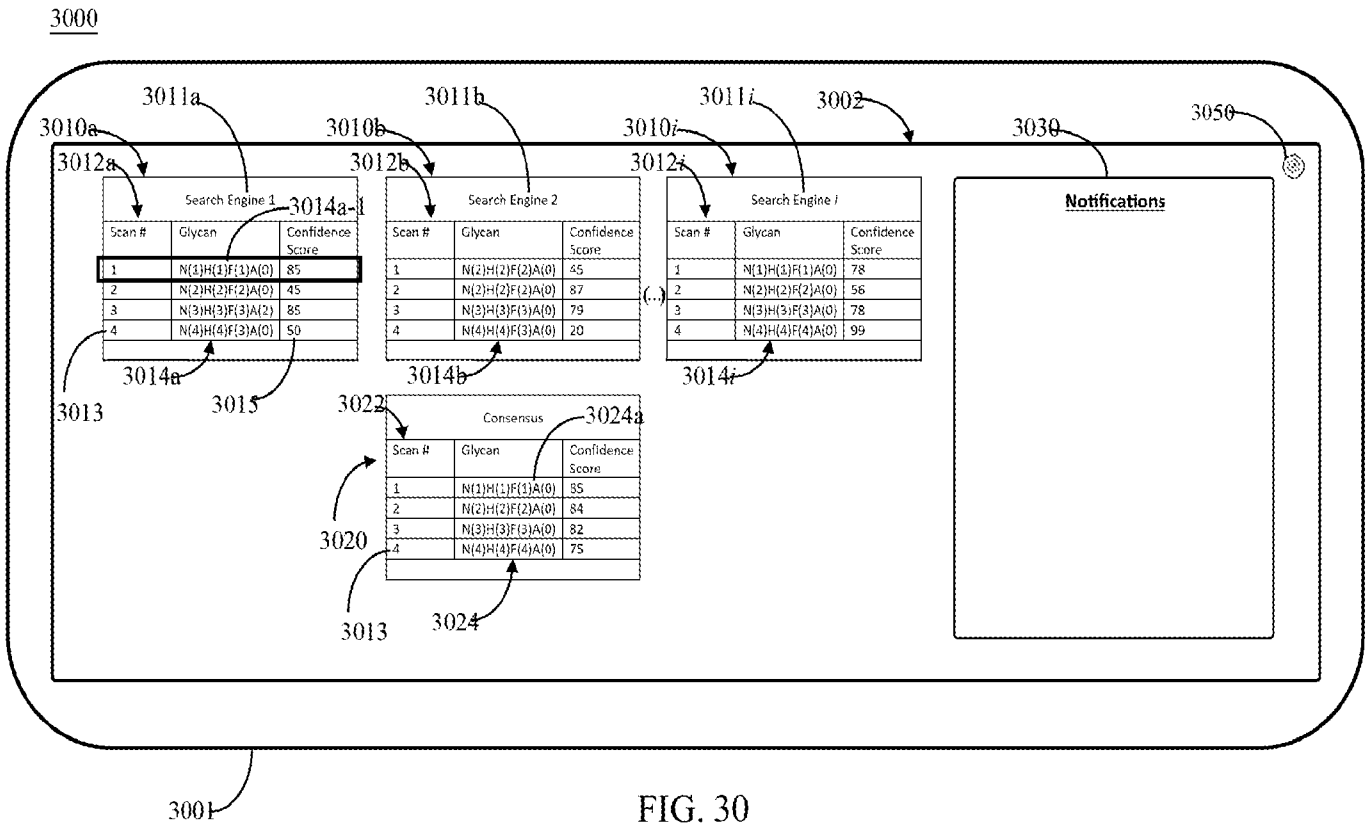

3000
3011a
3010a
3012a
3011b
3010b
3012b
3011i
3010i
3012i
3002
3030
3050
Search Engine 1
Scan # Glycan Confidence Score
1 N(1)H(1)F(1)A(0) 85
2 N(2)H(2)F(2)A(0) 45
3 N(3)H(3)F(3)A(2) 85
4 N(4)H(4)F(3)A(0) 50
3014a-1
3014a
3015
3013
Search Engine 2
Scan # Glycan Confidence Score
1 N(2)H(2)F(2)A(0) 45
2 N(2)H(2)F(2)A(0) 87
3 N(3)H(3)F(3)A(0) 79
4 N(4)H(4)F(3)A(0) 20
3014b
(...)
Search Engine i
Scan # Glycan Confidence Score
1 N(1)H(1)F(1)A(0) 78
2 N(2)H(2)F(2)A(0) 56
3 N(3)H(3)F(3)A(0) 78
4 N(4)H(4)F(4)A(0) 99
3014i
Notifications
3022
Consensus
3024a
Scan # Glycan Confidence Score
1 N(1)H(1)F(1)A(0) 85
2 N(2)H(2)F(2)A(0) 84
3 N(3)H(3)F(3)A(0) 82
4 N(4)H(4)F(4)A(0) 75
3020
3013
3024
3001

FIG. 30

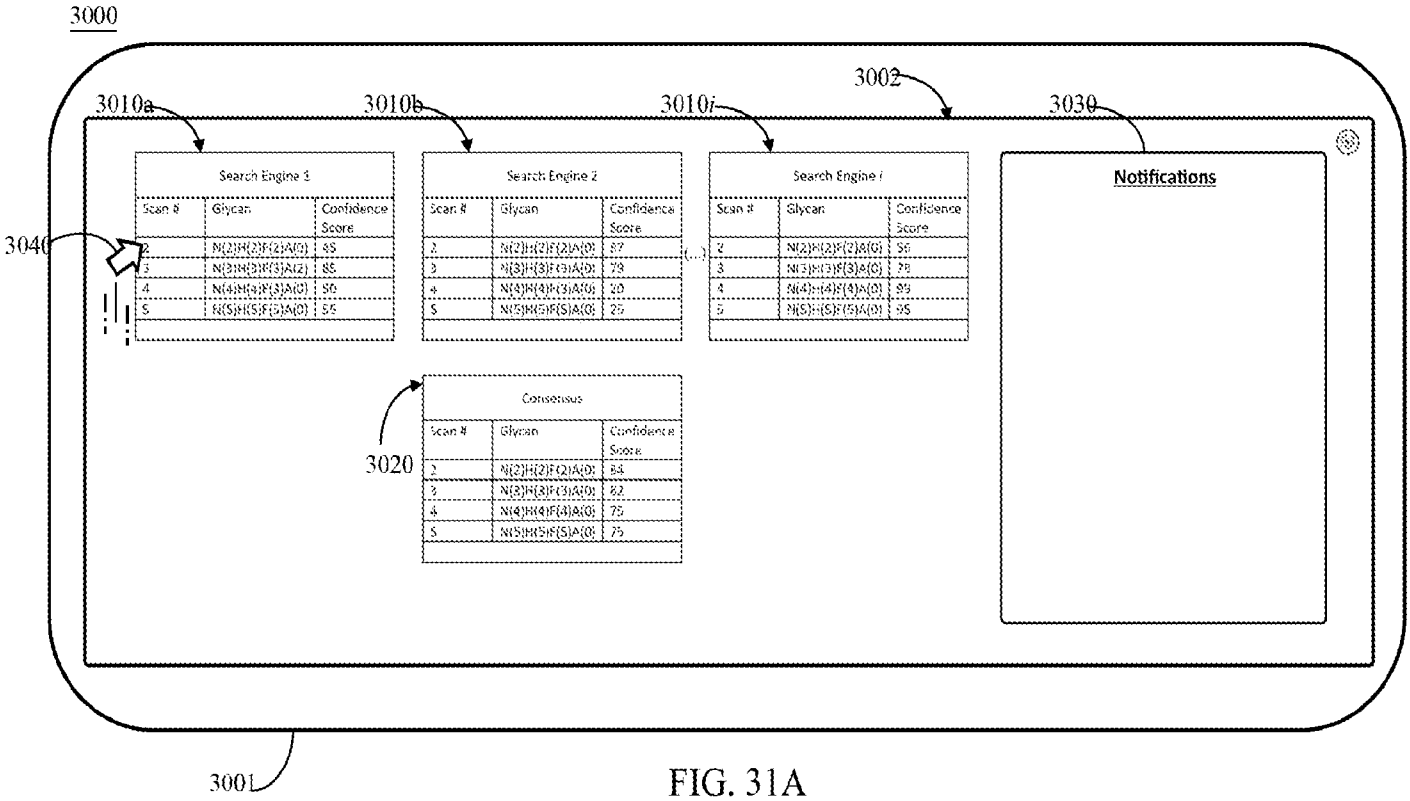

3000
3002
3010a
3010b
3010i
3030
Search Engine 1
Search Engine 2
Search Engine i
Notifications
Scan #
Glycan
Confidence Score
3040
Consensus
3020
3001

FIG. 31A

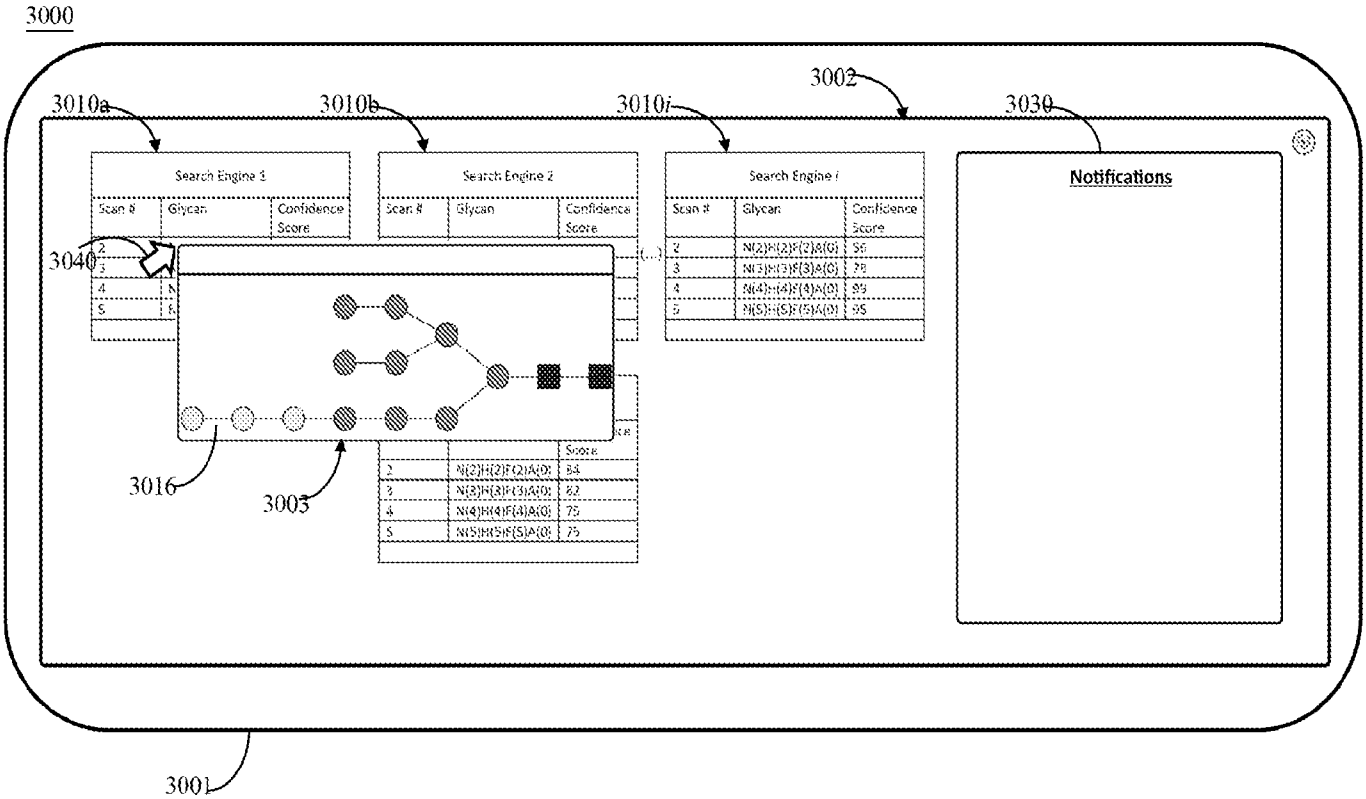

3000
3002
3010a
3010b
3010i
3030
Search Engine 1
Search Engine 2
Search Engine i
Scan #
Glycan
Confidence Score
Notifications
3040
3016
3003
3001

FIG. 31B

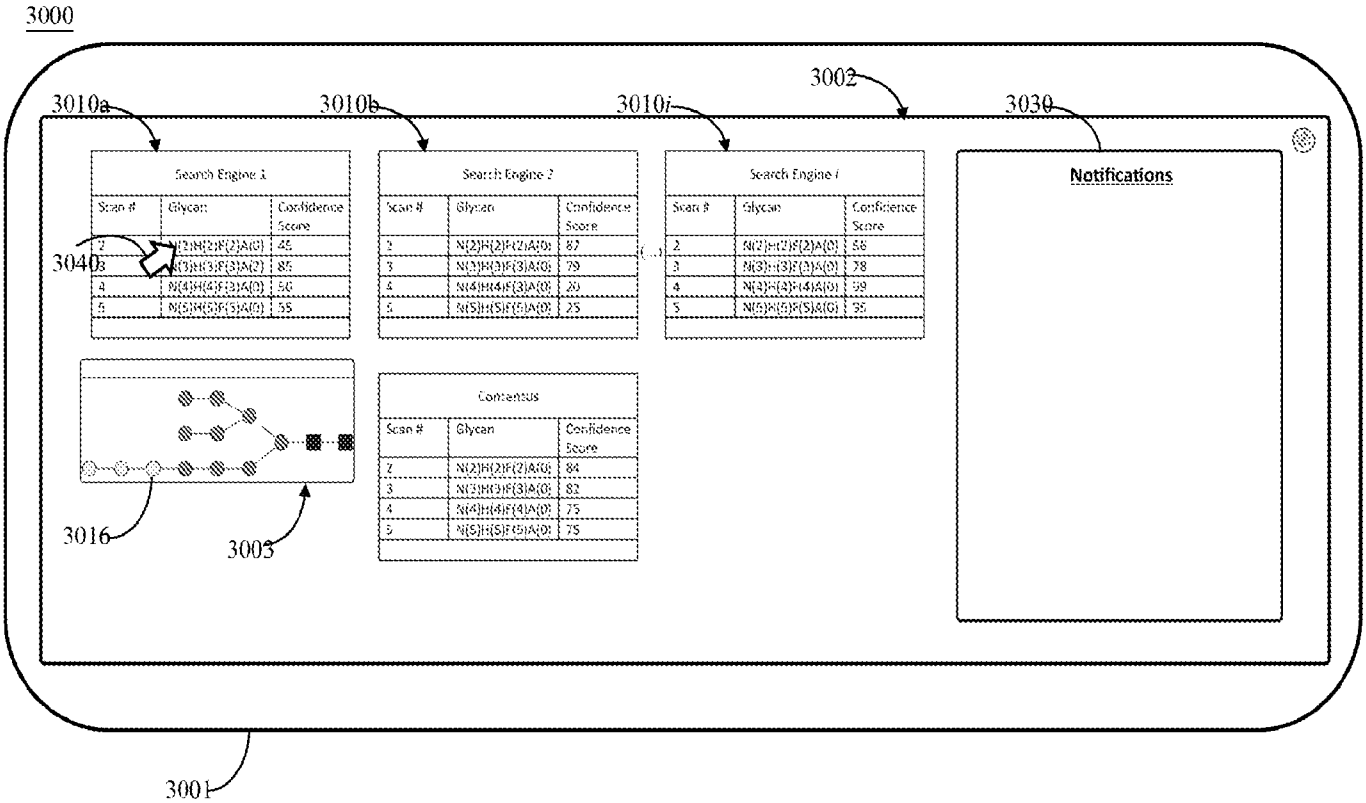

3000
3002
3010a
3010b
3010i
3030
Search Engine 1
Search Engine 2
Search Engine i
Notifications
Scan #
Glycan
Confidence Score
3040
Consensus
3016
3005
3001

FIG. 31C

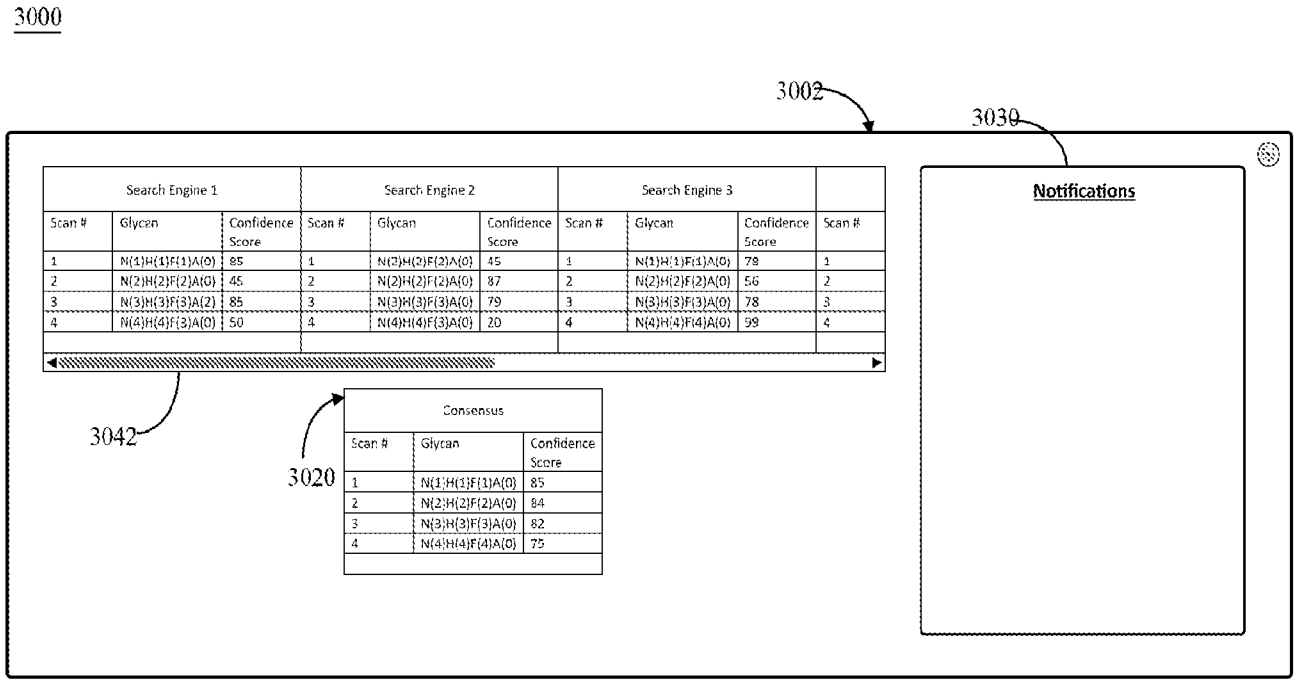

3000
3002
3030
Search Engine 1
Search Engine 2
Search Engine 3
Scan # | Glycan | Confidence Score | Scan # | Glycan | Confidence Score | Scan # | Glycan | Confidence Score | Scan #
1 | N(1)H(1)F(1)A(0) | 85 | 1 | N(2)H(2)F(2)A(0) | 45 | 1 | N(1)H(1)F(1)A(0) | 78 | 1
2 | N(2)H(2)F(2)A(0) | 45 | 2 | N(2)H(2)F(2)A(0) | 87 | 2 | N(2)H(2)F(2)A(0) | 56 | 2
3 | N(3)H(3)F(3)A(2) | 85 | 3 | N(3)H(3)F(3)A(0) | 79 | 3 | N(3)H(3)F(3)A(0) | 78 | 3
4 | N(4)H(4)F(3)A(0) | 50 | 4 | N(4)H(4)F(3)A(0) | 20 | 4 | N(4)H(4)F(4)A(0) | 99 | 4
Notifications
3042
3020
Consensus
Scan # | Glycan | Confidence Score
1 | N(1)H(1)F(1)A(0) | 85
2 | N(2)H(2)F(2)A(0) | 84
3 | N(3)H(3)F(3)A(0) | 82
4 | N(4)H(4)F(4)A(0) | 75

FIG. 32

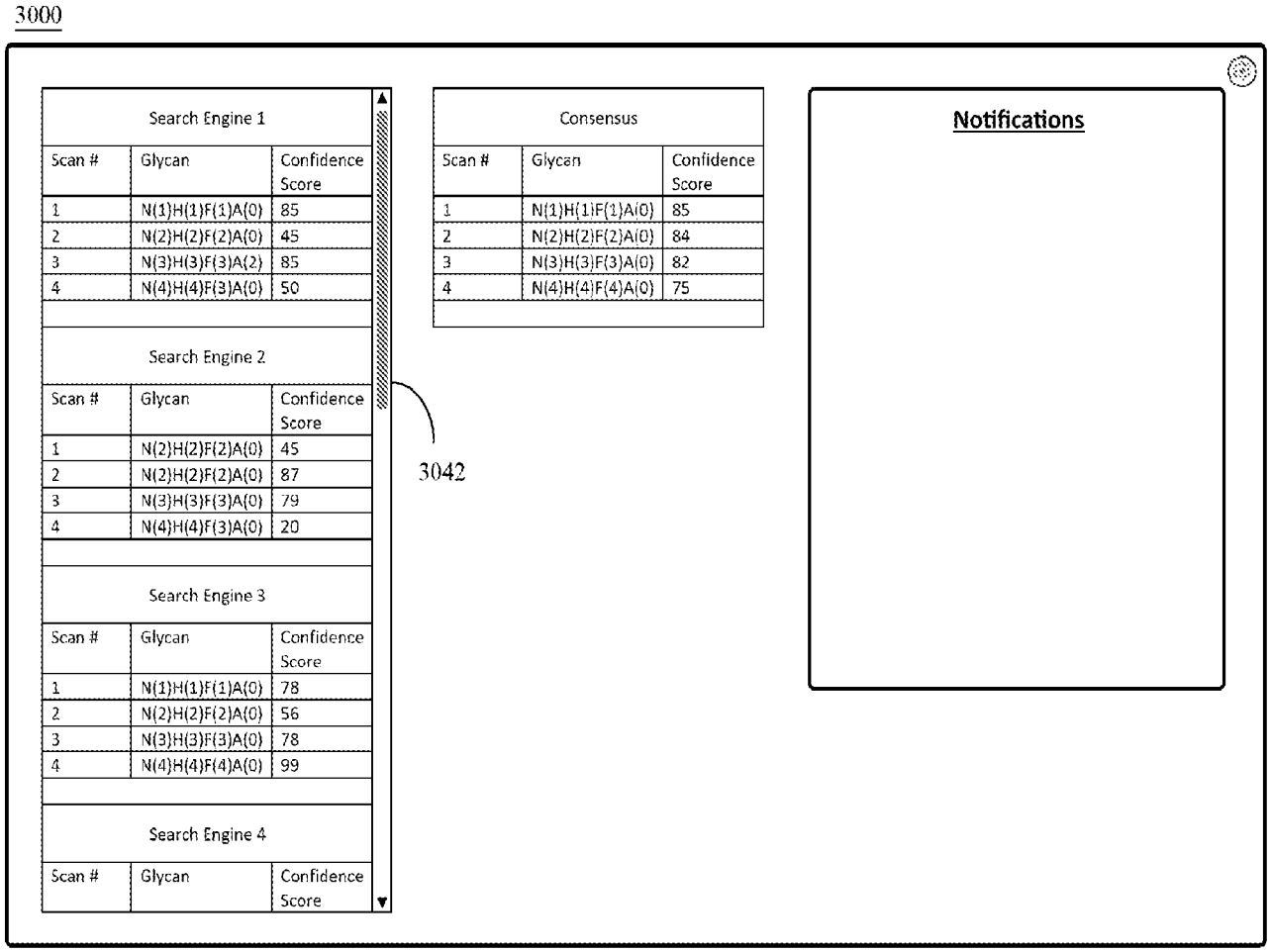

3000
Search Engine 1
Scan # | Glycan | Confidence Score
1 | N(1)H(1)F(1)A(0) | 85
2 | N(2)H(2)F(2)A(0) | 45
3 | N(3)H(3)F(3)A(2) | 85
4 | N(4)H(4)F(3)A(0) | 50
Search Engine 2
Scan # | Glycan | Confidence Score
1 | N(2)H(2)F(2)A(0) | 45
2 | N(2)H(2)F(2)A(0) | 87
3 | N(3)H(3)F(3)A(0) | 79
4 | N(4)H(4)F(3)A(0) | 20
Search Engine 3
Scan # | Glycan | Confidence Score
1 | N(1)H(1)F(1)A(0) | 78
2 | N(2)H(2)F(2)A(0) | 56
3 | N(3)H(3)F(3)A(0) | 78
4 | N(4)H(4)F(4)A(0) | 99
Search Engine 4
Scan # | Glycan | Confidence Score
3042
Consensus
Scan # | Glycan | Confidence Score
1 | N(1)H(1)F(1)A(0) | 85
2 | N(2)H(2)F(2)A(0) | 84
3 | N(3)H(3)F(3)A(0) | 82
4 | N(4)H(4)F(4)A(0) | 75
Notifications

FIG. 33

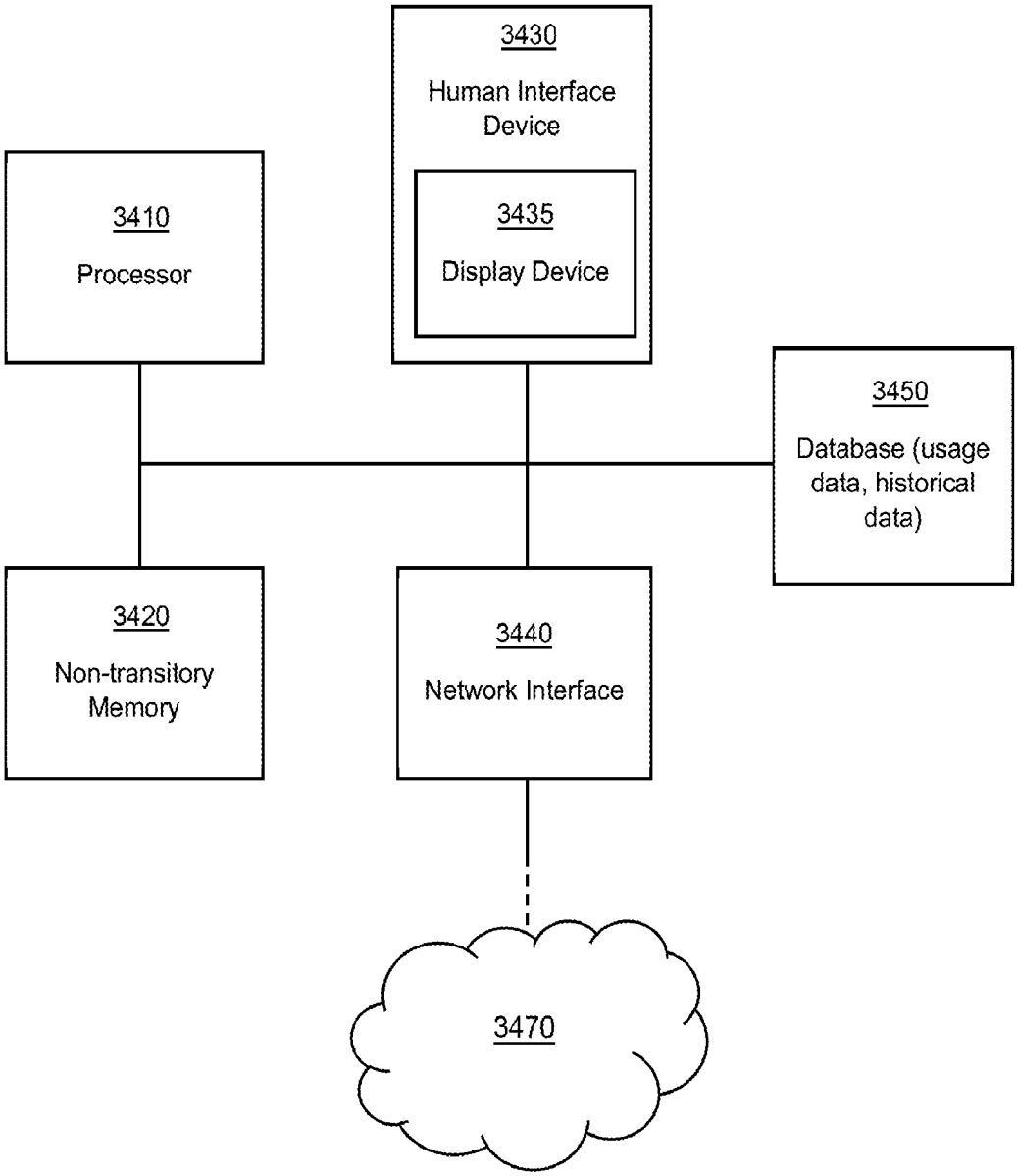

3430
Human Interface Device
3435
Display Device
3410
Processor
3450
Database (usage data, historical data)
3420
Non-transitory Memory
3440
Network Interface
3470

FIG. 34

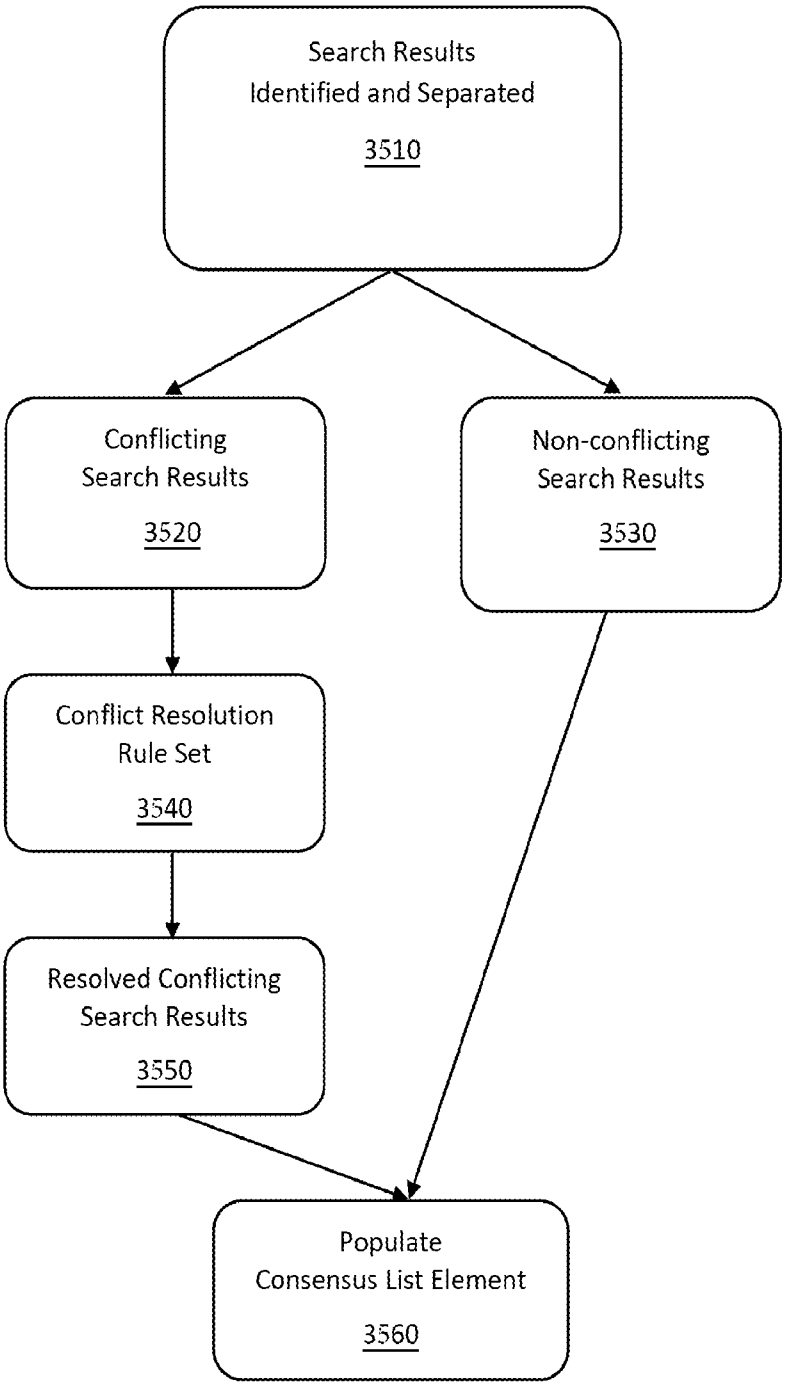

Search Results
Identified and Separated
3510
Conflicting
Search Results
3520
Non-conflicting
Search Results
3530
Conflict Resolution
Rule Set
3540
Resolved Conflicting
Search Results
3550
Populate
Consensus List Element
3560

FIG. 35

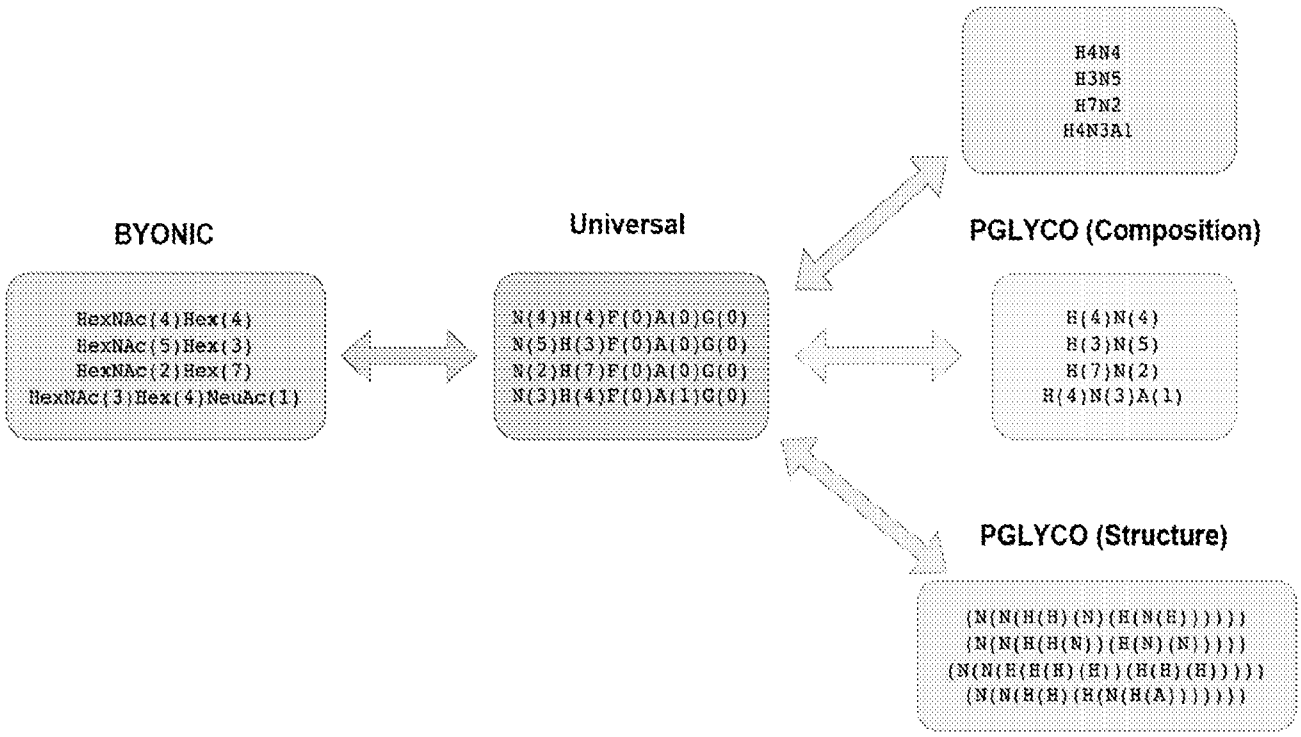

H4N4
H3N5
H7N2
H4N3A1
BYONIC
Universal
PGLYCO (Composition)
HexNAc(4)Hex(4)
HexNAc(5)Hex(3)
HexNAc(2)Hex(7)
HexNAc(3)Hex(4)NeuAc(1)
N(4)H(4)F(0)A(0)G(0)
N(5)H(3)F(0)A(0)G(0)
N(2)H(7)F(0)A(0)G(0)
N(3)H(4)F(1)A(1)G(0)
H(4)N(4)
H(3)N(5)
H(7)N(2)
H(4)N(3)A(1)
PGLYCO (Structure)
(N(N(H(H)(N)(H(N(H)))))
(N(N(H(H(N))(H(N)(N)))))
(N(N(H(H(H)(H))(H(H)(H)))))
(N(N(H(H)(H(N(H(A))))))

FIG. 36

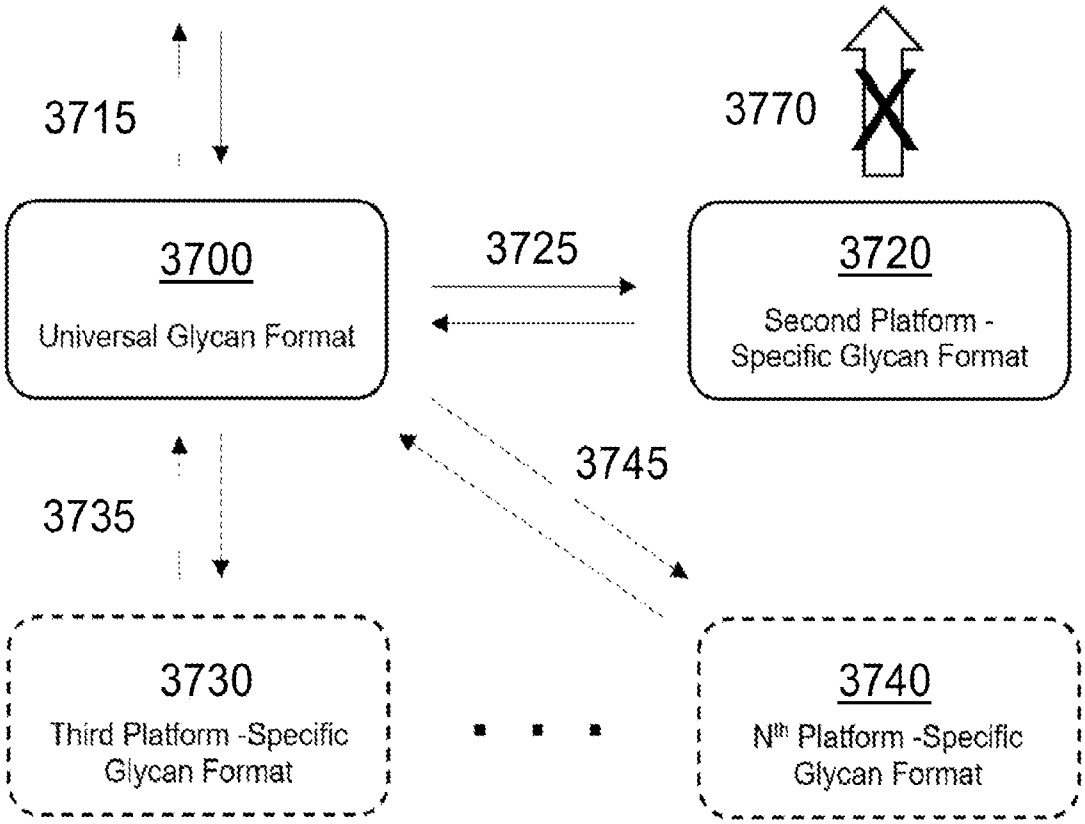

3715
3770
3700
Universal Glycan Format
3725
3720
Second Platform - Specific Glycan Format
3745
3735
3730
Third Platform -Specific Glycan Format
3740
Nth Platform -Specific Glycan Format

FIG. 37