TRAINING AND USE OF MACHINE-LEARNING MODELS FOR PREDICTING BIOLOGICAL CONDITIONS USING VOLATILE ORGANIC COMPOUNDS

RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application. No. 63/556,834 filed February 22, 2024, and U.S. Provisional Application No. 63/714,388 filed on October 31, 2024, each of which is hereby incorporated by reference in its entirety.

BACKGROUND

[0002] One method of disease diagnosis involves measuring molecular biomarkers from a sample from a subject and correlating those biomarkers with the disease. Typically, biomarkers are known molecular entities. For example, in certain cases, biomarkers can be protein biomarkers or nucleic acid biomarkers (e.g., DNA biomarkers, RNA biomarkers, or DNA modification biomarkers). In other cases, the biomarkers can be metabolites or volatile organic compounds.

SUMMARY

[0003] The present disclosure provides, among other things, compositions and methods useful for the diagnosis, detection, and/or monitoring of biological conditions including, without limitation, pathological states such as cancer. As set forth herein, a variety of biological conditions can be detected, diagnosed, and/or monitored by measurement of volatile organic compounds (VOCs) in a sample of a subject (e.g., a urine sample). In some embodiments, the subject from which the sample is extracted is a human. The present disclosure includes techniques for analysis thereof, including without limitation measurement of VOCs, e.g., by chromatographic separation followed by spectrometric analysis. In some embodiments, VOCs in a urine sample can be measured by a technique including gas chromatography followed by mass spectrometry (GC-MS). Moreover, VOC measurements can be analyzed, e.g., as provided herein, for diagnosis of a biological condition such as cancer.

[0004] The present disclosure includes the recognition that analysis of VOC measurements provides a powerful tool for diagnosis of biological conditions including cancer. The present

disclosure provides that even further advantages can be achieved by application of certain analytical techniques, including qualitative analysis for the presence or absence of peaks of interest, and/or application of machine learning to identify useful markers with VOC data sets, e.g., based on analysis of reference or training sample sets. When applied individually, or together, approaches provided herein have the potential to yield unprecedented advantages for the detection of biological conditions including cancer.

[0005] Among the various embodiments disclosed herein, the present disclosure provides analysis of urine VOCs using un-targeted chromatographic and mass spectrometry techniques, whereby many VOCs can be analyzed from the same sample. Moreover, in various embodiments, mass spectrometry for analysis of urine sample VOCs can apply a wide detection range for diverse mass-to-charge ratios. As compared to traditional diagnostics in which only one or a few individual biomarker molecules are analyzed, an un-targeted analysis can provide greater diagnostic power. Yet a further advantage of certain presently disclosed embodiments is the use of an effective internal standard for relative quantification of VOC concentration. The present disclosure includes the recognition that use of an internal standard can provide remarkable improvements in diagnostic quality.

[0006] The present disclosure further includes that compositions and methods of the present disclosure, when used for diagnosis, detection, and/or monitoring of a disease, disorder, or condition, can be used to determine therapy. For example, in some embodiments, compositions and methods of the present disclosure can be used in guiding selection of a particular therapeutic regimen, e.g., for treatment of cancer. In some embodiments, methods and compositions of the present disclosure can provide earlier, more cost-efficient, more sensitive, more specific, and/or more robust detection of relevant conditions, such as cancer. Further still, by providing early and correct detection, improvements in efficiency of therapy and clinical outcomes can be achieved.

[0007] The present disclosure further recognizes that measurements of VOCs in a urine sample can be represented in any form known to those in the art. When VOCs are measured by chromatographic separation followed by spectrometric analysis (e.g., by GC-MS), VOC measurements can be represented by mass-to-charge (m/z) ratios, retention time, and signal intensity (e.g., abundance or absorbance). In some embodiments, VOC measurements can be presented as two-dimensional data. In some embodiments, VOC measurements can be presented

as three-dimensional data. In some embodiments, VOC measurements can be represented as relative data, e.g., as compared to a standard, such as an internal standard.

[0008] In some embodiments, VOC measurements can be represented as amounts or concentrations corresponding to one or more selected peaks. In some embodiments, VOC measurements can be represented as a qualitative determination of the presence or absence of one or more selected peaks, optionally by application of a threshold level. In various optional embodiments, one or more peaks can correspond to a known or predicted molecule (e.g., a VOC). In other embodiments, peaks are not associated with a known or predicted molecule (e.g., a VOC), and/or any such association is not required for or included in diagnostic analyses.

[0009] In some embodiments, a biological condition such as cancer can be detected, diagnosed, and/or monitored based on the qualitative and/or quantitative measurement of one or more urine sample VOC mass spectrometry peaks, optionally where the one or more peaks are known or predicted to correspond to particular VOCs. In some embodiments, a biological condition such as cancer can be detected, diagnosed, and/or monitored based on the relative concentration of one or more urine sample VOC mass spectrometry peaks, e.g., as measured relative to an internal standard, optionally where the one or more peaks are known or predicted to correspond to particular VOCs.

[00010] In some embodiments, a biological condition such as cancer can be detected, diagnosed, and/or monitored based on the qualitative and/or quantitative measurement of one or more features of a urine sample VOC mass spectrometry abundance graph. In some embodiments, a biological condition such as cancer can be detected, diagnosed, and/or monitored based on features of a urine sample VOC mass spectrometry abundance graph, as measured relative to an internal standard, optionally where the features are known or predicted to correspond to particular VOCs. Thus, provided herein are compositions, methods and systems for diagnosing biological conditions (e.g., pathological states, e.g., cancer) using volatile organic compounds from a sample from a subject.

[00011] Data can be analyzed according to any of a variety of means, including by machine learning whereby measured sample data are compared with a reference generated based on analysis of reference or training sample set representative of a disease or condition. In various

embodiments, diagnosis, detection, and/or monitoring of a biological condition such as cancer can be achieved using a classification model.

[00012] In some embodiments, compositions, methods and systems of the present disclosure can include use of machine learning to develop a model (*e.g.*, a classification model, a regression model, or *etc.*) that can predict the state (*e.g.*, a classification) of a provided sample based on large datasets. In particular, the large datasets can include abundance data from three-dimensional or higher dimensional sample analysis. In one embodiment, a dataset that is analyzed by a machine learning algorithm can include mass spectra generated from samples fractionated by a chromatographic method, such as gas chromatography. Such datasets can include thousands of mass spectra collected over different retention times of the sample on a gas chromatograph. In contrast to other methods, in various embodiments, the learning algorithm may not abstract the abundance data into particular molecular biomarkers and use those molecules as predictors. Accordingly, the classifier can use the raw abundance data of a subject sample in making the classification, without identifying molecular species from the data. That is, classification can be performed without data reduction.

[00013] Classification can include not only a prediction of a particular state, but also, a confidence score that the prediction is correct. The present disclosure includes the recognition, that in various embodiments, a confidence score for diagnosis of a selected biological condition can be calculated at multiple time points over the course of analysis of a sample (e.g., over the course of mass spectrometry, e.g., during GC-MS of a sample), based on data having been collected from the sample as of that point in time, to determine if sufficient data has been collected to produce a diagnosis with the target level of confidence (confidence score). If diagnosis can be provided with the target level of confidence, the analysis can be regarded as complete and the analysis terminated, conserving resources (e.g., use of machines, reagents, and/or labor).

[00014] Accordingly, in certain embodiments, this confidence score can be used in determining how much data from a subject sample to collect. Mass spectra are collected as a function of retention time of samples eluting from the gas chromatograph. In some embodiments, compositions, methods and systems can apply a classifier to a first batch of data coming off the mass spectrometer and execute the classifier on that data. If the prediction has a

confidence score at or above a certain threshold, data collection can cease, and the prediction and confidence score can be reported out. However, if the confidence score is below a certain threshold, data collection can continue to expand the dataset, and classification can be performed on the expanded dataset. Data collection can continue in iterative fashion until the confidence score reaches the threshold or plateaus.

DEFINITIONS

[00015] ***About***: As used herein, term "about", when used in reference to a value, refers to a value that is similar, in context to the referenced value. In general, those skilled in the art, familiar with the context, will appreciate the relevant degree of variance encompassed by "about" in that context. For example, in some embodiments, the term "about" may encompass a range of values that within 25%, 20%, 19%, 18%, 17%, 16%, 15%, 14%, 13%, 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, or less of the referenced value.

[00016] ***Administration***: As used herein, the term "administration" typically refers to administration of a composition to a subject or system to achieve delivery of an agent that is, or is included in, the composition.

[00017] ***Agent***: As used herein, the term "agent" may refer to any chemical entity, including without limitation any of one or more of an atom, molecule, compound, amino acid, polypeptide, nucleotide, nucleic acid, protein, protein complex, liquid, solution, saccharide, polysaccharide, lipid, or combination or complex thereof.

[00018] ***Associated with***: Two events or entities are "associated" with one another, as that term is used herein, if the presence, level and/or form of one is correlated with that of the other. For example, a particular entity (e.g., compound, polypeptide, genetic signature, metabolite, microbe, etc.) is considered to be associated with a particular disease, disorder, or condition, if its presence, level and/or form correlates with incidence of and/or susceptibility to the disease, disorder, or condition (e.g., across a relevant population). In some embodiments, two or more entities are physically "associated" with one another if they interact, directly or indirectly, so that they are and/or remain in physical proximity with one another. In some embodiments, two or more entities that are physically associated with one another are covalently linked to one another; in some embodiments, two or more entities that are physically associated with one another are

not covalently linked to one another but are non-covalently associated, for example by means of hydrogen bonds, van der Waals interaction, hydrophobic interactions, magnetism, or a combination thereof.

[00019] ***Biomarker***: As used herein, the term "biomarker," consistent with its use in the art, refers to a to an entity, condition, or activity whose presence, level, or form, correlates with a particular biological event or state of interest, so that it is considered to be a "marker" of that event or state. To give but a few examples of biomarkers, in some embodiments, a biomarker can be or include a marker for a particular disease, disorder or condition, or can be a marker for qualitative or quantitative probability that a particular disease, disorder or condition can develop, occur, or reoccur, e.g., in a subject. In some embodiments, a biomarker can be or include a marker for a particular therapeutic outcome, or qualitative or quantitative probability thereof. Thus, in various embodiments, a biomarker can be predictive, prognostic, and/or diagnostic, of the relevant biological event or state of interest. In various embodiments, a biomarker can be an entity of any chemical class. For example, in some embodiments, a biomarker can be or include a VOC, a nucleic acid, a polypeptide, a lipid, a carbohydrate, a small molecule, an inorganic agent (e.g., a metal or ion), or a combination thereof. In some embodiments, a biomarker is a signal in data resulting of analysis of a sample, e.g., a peak in data generated by analysis of a sample using mass spectrometry, e.g., chromatography followed by mass spectrometry, e.g., GC-MS. Those of skill in the art will appreciate that a biomarker may be individually determinative of a particular biological event or state of interest, or may represent or contribute to a determination of the statistical probability of a particular biological event or state of interest. Those of skill in the art will appreciate that markers may differ in their specificity and/or sensitivity as related to a particular biological event or state of interest. In some embodiments, a biomarker is highly specific in that it reflects a high probability of a particular status of the biological event or state of interest. In some instances, there is an inverse relationship between specificity and sensitivity, such that increased specificity can come at the cost of sensitivity, or such that increased sensitivity can come at the cost of specificity. Those of skill in the art will appreciate that a useful biomarker need not have 100% specificity and/or 100% accuracy, and may reflect a balance of these or other considerations. In some instances, a biomarker can be referred to as a "marker."

[00020] ***Cancer***: As used herein, the term "cancer" refers to a disease, disorder, or condition in which cells exhibit relatively abnormal, uncontrolled, and/or autonomous growth, so that they display an abnormally elevated proliferation rate and/or aberrant growth phenotype characterized by a significant loss of control of cell proliferation. In some embodiments, a cancer can include one or more tumors. In some embodiments, a cancer can be or include cells that are precancerous (e.g., benign), malignant, pre-metastatic, metastatic, and/or non-metastatic. In some embodiments, a cancer can be or include a solid tumor. In some embodiments, a cancer can be or include a hematologic tumor. In some embodiments, cancer can be or include cells that are determined to have presence of cancer based on the prediction of the disclosed model, which may not be classified as precancerous by other diagnostic or clinical standards.

[00021] ***Chemotherapeutic agent***: As used herein, the term "chemotherapeutic agent," consistent with its use in the art, refers to one or more agents known, or having characteristics known to, treat or contribute to the treatment of cancer. In particular, chemotherapeutic agents include pro-apoptotic, cytostatic, and/or cytotoxic agents. In some embodiments, a chemotherapeutic agent can be or include alkylating agents, anthracyclines, cytoskeletal disruptors (e.g. microtubule targeting moieties such as taxanes, maytansine, and analogs thereof, of), epothilones, histone deacetylase inhibitors HDACs), topoisomerase inhibitors (e.g., inhibitors of topoisomerase I and/or topoisomerase II), kinase inhibitors, nucleotide analogs or nucleotide precursor analogs, peptide antibiotics, platinum-based agents, retinoids, vinca alkaloids, and/or analogs that share a relevant anti-proliferative activity. In some particular embodiments, a chemotherapeutic agent can include one or more of Actinomycin, All-trans retinoic acid, an Auiristatin, Azacitidine, Azathioprine, Bleomycin, Bortezomib, Carboplatin, Capecitabine, Cisplatin, Chlorambucil, Cyclophosphamide, Curcumin, Cytarabine, Daunorubicin, Docetaxel, Doxifluridine, Doxorubicin, Epirubicin, Epothilone, Etoposide, Fluorouracil, Gemcitabine, Hydroxyurea, Idarubicin, Imatinib, Irinotecan, Maytansine and/or analogs thereof (e.g. DM1) Mechlorethamine, Mercaptopurine, Methotrexate, Mitoxantrone, a Maytansinoid, Oxaliplatin, Paclitaxel, Pemetrexed, Teniposide, Tioguanine, Topotecan, Valrubicin, Vinblastine, Vincristine, Vindesine, Vinorelbine, or a combination thereof. In some embodiments, a chemotherapeutic agent can be utilized in the context of an antibody-drug conjugate. In some embodiments, a chemotherapeutic agent is one found in an antibody-drug

conjugate selected from hLL1-doxorubicin, hRS7-SN-38, hMN-14-SN-38, hLL2-SN-38, hA20-SN-38, hPAM4-SN-38, hLL1-SN-38, hRS7-Pro-2-P-Dox, hMN-14-Pro-2-P-Dox, hLL2-Pro-2-P-Dox, hA20-Pro-2-P-Dox, hPAM4-Pro-2-P-Dox, hLL1-Pro-2-P-Dox, P4/D10-doxorubicin, gemtuzumab ozogamicin, brentuximab vedotin, trastuzumab emtansine, inotuzumab ozogamicin, glembatumomab vedotin, SAR3419, SAR566658, BIIB015, BT062, SGN-75, SGN-CD19A, AMG-172, AMG-595, BAY-94-9343, ASG-5ME, ASG-22ME, ASG-16M8F, MDX-1203, MLN-0264, anti-PSMA ADC, RG-7450, RG-7458, RG-7593, RG-7596, RG-7598, RG-7599, RG-7600, RG-7636, ABT-414, IMGN-853, IMGN-529, vorsetuzumab mafodotin, and lorvotuzumab mertansine. In some embodiments, a chemotherapeutic agent can include one or more of farnesyl-thiosalicylic acid (FTS), 4-(4-Chloro-2-methylphenoxy)-N-hydroxybutanamide (CMH), estradiol (E2), tetramethoxystilbene (TMS), δ-tocatrienol, salinomycin, or curcumin.

[00022] ***Diagnosis***: As used herein, the term "diagnosis" refers to determining whether, and/or the qualitative or quantitative probability that, a subject has or will develop a disease, disorder, condition, or state. For example, in diagnosis of cancer, diagnosis can include a determination regarding the risk, type, stage, malignancy, or other classification of a cancer. In some instances, a diagnosis can be or include a determination relating to prognosis and/or likely response to one or more general or particular therapeutic agents or regimens.

[00023] ***Improve, increase, inhibit, decrease or reduce***: As used herein, the terms "improve", "increase", "inhibit", "decrease" and "reduce", and grammatical equivalents thereof, indicate qualitative or quantitative difference from a reference.

[00024] ***Level***: As used herein with respect to a molecule (e.g., a volatile organic compound), "level" is used to refer to a measure indicative of an amount, concentration, ratio, or activity of the molecule, e.g., in a particular context such as a tissue, sample, organism, or a context representative thereof. An amount can be, for example, a mass or number of molecules. A concentration can be an amount relative to a context value, e.g., per a unit of mass or volume. A ratio can be a relationship between two values, such as an experimental value and a reference control value. Activity can be a measure of a function associated with a molecule, and can in various instances be measured relative to a context value, e.g., per a unit of mass or volume. Those of skill in the art will appreciate that the metric by which a level is expressed can vary depending, e.g., on the assay and purpose. Those of skill in the art will further appreciate that

metrics such as amount, concentration, ratio, and activity are often interrelated and/or qualitatively or quantitatively informative of each other.

[00025] ***Reference***: As used herein, "reference" refers to a standard or control relative to which a comparison is performed. For example, in some embodiments, an agent, sample, sequence, subject, animal, or individual, or population thereof, or a measure or characteristic representative thereof, is compared with a reference, an agent, sample, sequence, subject, animal, or individual, or population thereof, or a measure or characteristic representative thereof. In some embodiments, a reference is a measured value. In some embodiments, a reference is an established standard or expected value. In some embodiments, a reference is a historical reference. A reference can be quantitative or qualitative. Typically, as would be understood by those of skill in the art, a reference and the value to which it is compared represents measure under comparable conditions (e.g., samples analyzed by comparable analytic techniques). Those of skill in the art will appreciate when sufficient similarities are present to justify reliance on and/or comparison. In some embodiments, an appropriate reference may be an agent, sample, sequence, subject, animal, or individual, or population thereof, under conditions those of skill in the art will recognize as comparable, e.g., for the purpose of assessing one or more particular variables (e.g., presence or absence of an agent or condition), or a measure or characteristic representative thereof. The terms diagnosis, detection, and/or monitoring of a disease, disorder, or condition can be used interchangeably, where detection can refer to, e.g., the determining of the presence or absence of a disease, disorder, or condition; diagnosis includes detection in a subject; and monitoring includes detection/diagnosis to track the presence, absence, or progress of a disease, disorder, or condition, optionally applying an assay to the same sample or subject at multiple points in time.

[00026] ***Sample:*** As used herein, the term "sample" typically refers to an aliquot of material obtained or derived from a source of interest (e.g., a sample from urine of a subject). In some embodiments, a sample is a "primary sample" obtained directly from a source of interest (e.g., a sample composed of urine). In some embodiments, as will be clear from context, the term "sample" refers to a preparation that is obtained by processing of a primary sample (e.g., by removing one or more components of and/or by adding one or more agents to a primary sample,

e.g., a sample produced by or during purification of VOCs from a sample of urine and/or a sample prepared for chromatography).

[00027] ***Sensitivity***: As used herein, the "sensitivity" of a diagnostic test refers to the percentage of samples that are characterized by the presence of the event or state of interest for which measurement of the diagnostic test accurately indicates presence of the event or state of interest (true positive rate). In various embodiments, characterization of the positive samples is independent of the diagnostic test, and can be achieved by any relevant measure, e.g., any relevant measure known to those of skill in the art. Thus, sensitivity reflects the probability that a diagnostic test would detect the presence of the event or state of interest when measured in a sample characterized by presence of that event or state of interest. In particular embodiments in which the event or state of interest is cancer, sensitivity refers to the probability that a diagnostic test would detect the presence of cancer in a subject that has cancer.

[00028] ***Specificity***: As used herein, the "specificity" of a diagnostic test refers to the percentage of samples that are characterized by absence of the event or state of interest for which measurement of the diagnostic test accurately indicates absence of the event or state of interest (true negative rate). In various embodiments, characterization of the negative samples is independent of the diagnostic test, and can be achieved by any relevant measure, e.g., any relevant measure known to those of skill in the art. Thus, specificity reflects the probability that the diagnostic test would detect the absence of the event or state of interest when measured in a sample not characterized that event or state of interest. In particular embodiments in which the event or state of interest is cancer, specificity refers to the probability that a diagnostic test would detect the absence of cancer in a subject lacking colorectal cancer.

[00029] ***Subject***: As used herein, the term "subject" refers to an organism, such as a human. In some embodiments, a subject is suffering from a disease, disorder or condition. In some embodiments, a subject is susceptible to a disease, disorder, or condition. In some embodiments, a subject displays one or more symptoms or characteristics of a disease, disorder or condition. In some embodiments, a subject is not suffering from a disease, disorder or condition. In some embodiments, a subject does not display any symptom or characteristic of a disease, disorder, or condition. In some embodiments, a subject has one or more features characteristic of susceptibility to or risk of a disease, disorder, or condition. In some embodiments, a subject is a

subject that has been tested for a disease, disorder, or condition, and/or to whom therapy has been administered. In some instances, a human subject can be interchangeably referred to as a "patient" or "individual." A subject administered an agent associated with treatment of a disease, disorder, or condition with which the subject is associated can be referred to as a subject in need of the agent, i.e., as a subject in need thereof.

[00030] ***Therapeutic agent:*** As used herein, the term "therapeutic agent" refers to any agent that elicits a desired pharmacological effect when administered to a subject. In some embodiments, an agent is considered to be a therapeutic agent if it demonstrates a statistically significant effect across an appropriate population. In some embodiments, the appropriate population can be a population of model organisms or a human population. In some embodiments, an appropriate population can be defined by various criteria, such as a certain age group, gender, genetic background, preexisting clinical conditions, etc. In some embodiments, a therapeutic agent is a substance that can be used for treatment of a disease, disorder, or condition. In some embodiments, a therapeutic agent is an agent that has been or is required to be approved by a government agency before it can be marketed for administration to humans. In some embodiments, a therapeutic agent is an agent for which a medical prescription is required for administration to humans.

[00031] ***Treatment***: As used herein, the term "treatment" (also "treat" or "treating") refers to administration of a therapy that partially or completely alleviates, ameliorates, relieves, inhibits, delays onset of, reduces severity of, and/or reduces incidence of one or more symptoms, features, and/or causes of a particular disease, disorder, or condition, or is administered for the purpose of achieving any such result. In some embodiments, such treatment can be of a subject who does not exhibit signs of the relevant disease, disorder, or condition and/or of a subject who exhibits only early signs of the disease, disorder, or condition. Alternatively or additionally, such treatment can be of a subject who exhibits one or more established signs of the relevant disease, disorder and/or condition. In some embodiments, treatment can be of a subject who has been diagnosed as suffering from the relevant disease, disorder, and/or condition. In some embodiments, treatment can be of a subject known to have one or more susceptibility factors that are statistically correlated with increased risk of development of the relevant disease, disorder, or condition.

BRIEF DESCRIPTION OF THE DRAWINGS

[00032] The accompanying drawings, which are incorporated herein and form a part of the specification, illustrate exemplary embodiments and, together with the description, further serve to enable a person skilled in the pertinent art to make and use these embodiments and others that will be apparent to those skilled in the art. The invention will be more particularly described in conjunction with the following drawings wherein:

[00033] **Figure 1** shows an exemplary methodology for performing statistical analysis on raw data generated from volatiles in a urine sample. VOCs from the sample are captured using stir bar sorptive extraction. The extracted samples are analyzed by GC-MS to produce a raw data set. In certain embodiments, the raw data set is subject to machine learning. Correlation between data from the data set and a pathologic state, such as bladder cancer, is determined.

[00034] **Figure 2** shows an exemplary GC chromatogram showing analyte abundance is a function of retention time.

[00035] **Figure 3** shows an exemplary mass spectrum at one retention time. Signal intensity is a function of mass-to-charge ratio.

[00036] **Figure 4** shows an exemplary raw abundance matrix of data from a GC-MS experiment in which each feature represents abundance of an analyte as a function of retention time and mass-to-charge ratio.

[00037] **Figure 5** shows an abundance matrix from a GC-MS experiment. The x-axis represents mass-to-charge ratio. The Y axis represents retention time with zero retention time at the top. Darkness of color represents intensity of signal.

[00038] **Figure 6A** is a block diagram illustrating a system for generating disease prediction(s) from mass spectrometry reads according to embodiments of the present disclosure.

[00039] **Figure 6B** is a diagram illustrating a system for generating disease prediction(s) from mass spectrometry reads according to embodiments of the present disclosure.

[00040] **Figure 6C** is a block diagram illustrating a system for generating disease prediction(s) from mass spectrometry reads according to embodiments of the present disclosure.

[00041] **Figure 7A** is a block diagram illustrating a system for training a machine learning model to generate disease prediction(s) from a training set of mass spectrometry reads and subject disease data according to embodiments of the present disclosure.

[00042] **Figure 7B** is a block diagram illustrating a system for training a machine learning model to generate disease prediction(s) from a training set of abundance matrixes**.**

[00043] **Figure 7C** is a block diagram illustrating a system for training a machine learning model to generate disease prediction(s) from a training set of mass spectrometry reads and subject disease data according to embodiments of the present disclosure.

[00044] **Figures 8A** and **8B** are user interfaces illustrating example outputs of a dashboard according to embodiments of the present disclosure.

[00045] **Figure 9** is a user interface illustrating a dashboard according to embodiments of the present disclosure.

[00046] **Figure 10A** is a flowchart illustrating a process of generating a disease prediction according to embodiments of the present disclosure.

[00047] **Figure 10B** is a flowchart illustrating a process of generating a disease prediction according to embodiments of the present disclosure.

[00048] **Figures 11A** is a flowchart illustrating a method of training a machine-learning model according to embodiments of the present disclosure.

[00049] **Figures 11B** is a flowchart illustrating a method of training a machine-learning model according to embodiments of the present disclosure.

[00050] **Figure 12** illustrates a Receiver Operated Curve (ROC) having an area under the curve (AUC) of 0.96 for a classification to diagnose prostate cancer. The classification algorithm uses raw features, but not identified molecular biomarkers, making the diagnosis.

[00051] **Figure 13** is a diagram illustrating a computing node according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

[00052] Cancer is one of the leading causes of death and is expected to affect millions annually. Treatment efficacy and clinical outcomes can be substantially improved when cancer is diagnosed at an early clinical stage, underpinning the importance of the timely and effective

diagnosis. Molecular diagnostic testing is particularly important for cancer at least in part because cancer is often difficult to detect, particularly at early stages, due to difficulty of diagnosis based on physiological symptoms alone. Accordingly, there remains a need for compositions and methods for the detection and/or diagnosis of biological conditions including cancers, e.g., with high sensitivity and specificity.

[00053] The present disclosure includes, among other things, compositions and methods useful for the detection, diagnosis, and/or monitoring of biological conditions (e.g., cancer) in a subject based on the measurement of VOCs of a sample from the subject. As provided herein, VOCs can be measured using chromatography followed by mass spectrometry (e.g., GC-MS), optionally together with forms of machine learning for analysis (e.g., classification). In some embodiments, the present disclosure includes measuring a plurality of VOCs of a sample, in some cases hundreds or thousands of VOCs. In some embodiments, the present disclosure includes un-targeted analysis of VOCs, e.g., where assay steps are engineered to capture a broad set of potential VOCs and many VOCs can be analyzed from the same sample. For example, in some embodiments, mass spectrometry for analysis of urine sample VOCs can apply a wide detection range for diverse mass-to-charge ratios. In some embodiments, analyses provided herein can employ relative quantification (*e.g.*, using an internal standard) of VOC concentrations.

[00054] Further still, the present disclosure provides that even further advantages can be achieved by application of certain analytical techniques, including qualitative analysis for the presence or absence of peaks of interest, and/or application of machine learning to identify useful markers with VOC data sets, e.g., based on analysis of reference or training sample sets.

[00055] The present disclosure further provides the recognition that the measurement and analysis of VOCs in urine provide surprising and advantageous utility in the detection, diagnosis, and/or monitoring of biological conditions, and that the compositions and methods disclosed herein can provide improved clinical outcomes when used in conjugation with suitable therapies for the biological conditions.

I. Samples and Sample Preparation

[00056] Volatile organic compounds (VOCs) are compounds having relatively low boiling point and high vapor pressures at room temperature. As a consequence, they readily evaporate at normal or standard temperatures and pressures, and can be emitted as gases from certain solids or liquids. Broadly, there are thousands of distinct VOCs. Many volatile compounds are small organic compounds. They can have molecular weights of up to about 200 g per mole. Examples of VOCs can include hydrocarbons, alcohols, aldehydes, organic acids. Some well-known examples are ethanol and acetone. VOCs can be present in a variety of biological samples, such as in exhaled breath or whole blood. VOCs present in a liquid sample have a greater presence in air adjacent to the sample than various other types of compounds present in the sample. As referred to herein, VOCs include both semi-volatile organic compounds (SVOCs) and very volatile organic compounds (VVOCs), although many VVOCs exist largely or exclusively as gases.

[00057] Samples used in the compositions, methods, and systems disclosed herein are those samples comprising VOCs. A sample can be a biological sample, which is a sample that includes material of biological origin. A sample or biological sample of the present disclosure can be collected with means known to those skilled in the art. In certain embodiments, a biological sample may be collected through non-invasive means (e.g., urination). In some embodiments, the biological sample may be collected through invasive means (e.g., biopsy). Exemplary biological samples from which volatile organic compounds can be analyzed with the compositions, methods and systems disclosed herein include, without limitation, urine, blood, saliva, breath, septum, breast milk, exhaled breath, intestinal gas or flatus, cerebrospinal fluid, mucus, lymph, feces, spinal fluid, peritoneal fluid, lymphatic fluid, synovial fluid, tears, seminal fluid, vaginal fluids, pulmonary effusion, and serosal fluid. In some embodiments, the biological sample is urine.

[00058] In various embodiments, a sample of the present disclosure, such as a urine sample, can be prepared for analysis, e.g., by chromatography followed by mass spectrometry, e.g., by GC-MS. In some embodiments, VOCs can be extracted from a sample to produce a VOC sample used in subsequent steps of analysis such as chromatography followed by mass spectrometry, e.g., by GC-MS.

[00059] One method of extracting VOCs from a sample, such as a urine sample, is to contact the sample with a substrate that adsorbs VOCs. Examples of substrates that absorb VOCs include metal organic frameworks (MOFs), activated carbons (ACs), hypercrosslinked polymeric resin (HPR), and zeolites. Mechanism of VOC adsorption can include one or more of include electrostatic attraction, interaction between polar VOCs and hydrophilic sites, interaction between non-polar VOCs and hydrophobic sites, and partition in non-carbonized portions. Adsorption capacity can be impacted by various factors and can, for example, increases with the specific surface area, pore volume, and/or presence of certain surface chemical functional groups, and/or with decreasing pore size.

[00060] One particular example of a substrate that adsorbs VOCs from a liquid sample such as urine is polydimethylsiloxane (PDMS), a sorbent that can be used, e.g., in Stir Bar Sorptive Extraction (SBSE). SBSE can include sorption of an analyte(s) onto a thick film of polydimethylsiloxane (PDMS) coated on the magnet of a stir bar, e.g., incorporated into a glass jacket. Analytes are sampled by introducing the stir bar directly into the liquid sample or suspending it in the matrix headspace for a fixed time. After sampling analytes can be recovered from the PDMS by thermal desorption and on-line analyzed by cGC or cGC–MS. Exemplary PDMS stir bars include a TWISTER® Stir Bar (GERSTEL®), which allows analysis of organic compounds from aqueous solutions. While stirring, it adsorbs and concentrates the VOCs onto its sorbent coating.

[00061] Samples of the present disclosure can be further prepared before analysis, e.g., by addition of reagents or materials that facilitate chromatography and spectrometry, optionally including addition of an internal standard to the sample. An internal standard, or reference standard, can refer to an agent that is added to a sample composition in a predetermined amount and/or concentration, whereby, during analysis of the sample, the measured value for the standard provides a benchmark against which the measured value for test analytes can be compared to determine the relative amount and/or concentration of the test analytes in the sample.

[00062] For example, in some embodiments, Mirex, or the like, can be added to a sample of the present disclosure as an internal standard. Mirex ($C_{10}Cl_{12}$; CAS: 2385-85-5; MW 545.54 KD) is an organochlorine insecticide (e.g., in PESTANAL®), used herein as an internal

reference standard for determination of the presence or absence, or relative amount or concentration, of VOCs (or of any given point on a 2-dimensional, 3-dimensional, or further multi-dimensional representation of data, e.g., from GC-MS) following chromatography and mass spectrometry (e.g., GC-MS). The present disclosure includes the unexpected recognition that inclusion of an internal standard is particularly advantages for relative determination of the presence or absence, or relative amount or concentration, of VOCs following chromatography and mass spectrometry (e.g., GC-MS), e.g., using Mirex as an internal standard.

[00063] The present disclosure further includes the finding that samples prepared in accordance with various embodiments of the present disclosure are highly stable, an important characteristic for clinical application of methods provided herein to diagnose biological conditions such as cancer. In some embodiments, a sample may be analyzed immediately after collection. In some embodiments, a sample may be stored before analysis. In some embodiments, a sample may be stored at a temperature below 0°C (e.g., -20°C or -80°C). In some embodiments, the sample may be stored for up to or at least about 5 years (e.g., one month, about two months, about three months, about four months, about five months, about six months, about seven months, about eight months, about nine months, about 10 months, about 11 months, about 1 year, about 1 1/2 years, about 2 years, about 3 years, about 4 years, or about 5 years) at a temperature below 0°C (e.g., -20°C or -80°C). In some embodiments, the sample is stable in storage for at least 1 year a temperature below 0°C (e.g., -20°C or -80°C). In some embodiments, the sample is stable in storage for at least 1 year at -80°C. In various embodiments, a sample is regarded as stable for at least an indicated period if application of the a particular diagnostic assay provides the same diagnostic result at the end of the indicated period as would have been found by the same assay prior to storage, e.g., e.g., immediately or shortly after collection (e.g., within 1 day of collection, within 2 days of collection, or within 3 days of collection).

II. Fractionation and Spectrometry

[00064] The present disclosure includes methods of measuring the presence or absence, or relative amount or concentration, of VOCs (or of any given point on a 2-dimensional, 3-

dimensional, or further multi-dimensional representation of data, e.g., from GC-MS) following fractionation (e.g., by chromatography) and mass spectrometry (e.g., GC-MS).

[00065] (A) Fractionation (e.g., Chromatography)

[00066] Analytes in a sample can be subject to a first fractionation method to separate analytes in the sample into different fractions. Such methods can include, without limitation, chromatography and electrophoresis. In separation methods in which analytes travel through a matrix, output can be a function of residence time or retention time. As such, data collected also can be assigned to particular residence times or retention times. Furthermore, data can be collected for all or part of the residence time or retention time it takes for complete or essentially complete processing of a sample by the fractionation method and/or for the final analytes to be separated in the run.

[00067] Chromatography is a method of separating analytes in a sample according to differences in one or more properties. Chromatography typically comprises a stationary phase and a mobile phase. The stationary phase typically comprises a solid support that interacts differently with analytes based on certain characteristics. The mobile phase can be a liquid or gas in which analytes are present. In various embodiments, the mobile phase is moved over the solid phase, and analytes can travel through the solid phase at a rate that reflects their degree of interaction with the solid phase. Samples including VOCs can be prepared for chromatography and applied to a chromatography system and/or substrate according to methods known to those of skill in the art.

[00068] In various embodiments, chromatography can be applied to detect the presence or absence, or relative amount or concentration, of VOCs (or of any given point on a 2-dimensional, 3-dimensional, or further multi-dimensional representation of data, e.g., from GC-MS). . Examples of chromatography include gas chromatography, liquid chromatography, high performance liquid chromatography, size exclusion chromatography, affinity chromatography, anion exchange chromatography, cation exchange chromatography, gel filtration chromatography, hydrophobic interaction chromatography, ion exchange chromatography, reverse phase chromatography, paper chromatography, and thin-layer chromatography. Chromatography can optionally be combined with, e.g., electrophoresis and/or mass spectrometry methods or systems.

[00069] In liquid chromatography (LC) stationary phase is a liquid absorbed to a solid support. The mobile phase comprises a liquid in which the analytes are dissolved. High-performance liquid chromatography (HPLC) uses high pressure to pass the mobile stage through the stationary phase.

[00070] In supercritical liquid chromatography, the sample is dissolved in a supercritical fluid, for example, carbon dioxide, and that is passed through a chromatographic column. Typically, the stationary phase is a silica-based material modified with polar functional groups.

[00071] In thin layer chromatography (TLC) stationary phase is a layer of material that coats a solid support such as a glass plate. The solid phase is a liquid that comprises the analyte.

[00072] In parallel comprehensive 2-dimensional gas chromatography, a sample is divided into aliquots in each aliquot is subject to a gas chromatography process that separates analytes based on different physical and chemical characteristics.

[00073] In series comprehensive 2-dimensional gas chromatography, a sample is subjected to first and then second gas chromatography processes in which each process separates analytes based on different physical and chemical characteristics.

[00074] The measurement of analytes separated by chromatography produces a chromatogram in which the abundance of an analyte is a function of its position in the separation stream. In the case of gas chromatography or liquid chromatography, abundance is a function of retention time, that is, the time it takes an analyte to pass through the solid phase (typically a column).

[00075] In gas chromatography (GC) the stationary phase is a thin layer of liquid or polymer coated onto a solid support. The mobile phase is a gas. Gas chromatography is typically used to analyze volatile compounds.

[00076] For gas chromatography, analytes can be collected from a sample (e.g., a urine sample) using, for example, stir bar sorptive extraction (SBSE). In this method, a stir bar that is affinity for certain analytes in a sample is introduced into the sample for analyte collection. In certain embodiments, analytes can be collected from a urine sample, e.g., by SBSE or solvent desorption prior to chromatography, e.g., by incubation of the sample with a stir bar coated with polydimethylsiloxane (e.g., a TWISTER® Stir Bar). In certain embodiments, analytes can be collected on different stir bars that attract analytes based on different properties.

[00077] In various embodiments in which VOCs are adsorbed by a substrate (e.g., in SBSE) prior to chromatography (e.g., gas chromatography), the VOCs are released from the substrate prior to chromatographic separation. In various embodiments, the process of releasing VOCs from a substrate by which they have been adsorbed is referred to as thermal desorption. Thermal desorption can include release of VOCs from the sorbent using heat and an inert gas flow. Once released, VOCs can be focused onto a secondary trap for re-concentration and rapid injection into a GC-MS. Because the inert gas flow rate is typically greater than the column flow rate, the secondary trap permits appropriate rate of flow through the chromatographic column. This process provides improved chromatography and very low detection limits.

[00078] **(B) Mass Spectrometry**

[00079] Analytes in a sample can be detected by mass spectrometry. In various embodiments, after fractionation (e.g., by chromatography, e.g. by gas chromatography), fractionated sample can be subjected to mass spectrometry. For instance, mass spectrometry can be coupled with separation techniques such as liquid chromatography or gas chromatography. In some embodiments, gas chromatography can be coupled with mass spectrometry (GC-MS).

[00080] Mass spectrometers typically include an ion source to ionize analytes, and one or more mass analyzers to determine mass. Mass analyzers can be used together in tandem mass spectrometers. Ionization methods include, among others, electrospray or laser desorption ionization. Mass analyzers include quadrupoles, ion traps, time-of-flight instruments and magnetic or electric sector instruments. In certain embodiments, the mass spectrometer is a tandem mass spectrometer (e.g., "MS-MS") that uses a first mass analyzer to select ions of a certain mass and a second mass analyzer to analyze the selected ions. One example of a tandem mass spectrometer is a triple quadrupole instrument, the first and third quadrupoles act as mass filters, and an intermediate quadrupole functions as a collision cell. Mass spectrometry also can be coupled with up-stream separation techniques, such as liquid chromatography or gas chromatography. So, for example, gas chromatography coupled with mass spectrometry can be referred to as "GC-MS".

[00081] The present disclosure includes that various mass spectrometry methods and systems known in the art can be used in accordance with the present disclosure, including without limitation mass spectrometry using various ionization systems (e.g., Electron Ionization,

Electrospray Ionization, or Matrix Assisted Laser Desorption), mass analyzers (e.g., Quadrupole, Ion Trap, Time-of-Flight (TOF), Magnetic sector, or Orbitrap), and detectors (e.g., electron multiplier, faraday cup, photomultiplier conversion dynode, or array detectors). In various embodiments mass spectrometry can be tandem mass spectrometry (e.g., MS/MS). Examples of mass spectrometry include, e.g., matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, electrospray ionization (ESI) mass spectrometry, surface-enhanced laser deorption/ionization-time of flight (SELDI-TOF) mass spectrometry, quadrupole-time of flight (Q-TOF) mass spectrometry, atmospheric pressure photoionization mass spectrometry (APPI-MS), Fourier transform mass spectrometry (FTMS), matrix-assisted laser desorption/ionization-Fourier transform-ion cyclotron resonance (MALDI-FT-ICR) mass spectrometry, or secondary ion mass spectrometry (SIMS).

[00082] Therefore, in various embodiments, provided herein are novel compositions, methods and systems for predicting pathological states through the analysis of biological samples and biofluids using advanced analytical techniques, such as GC-MS, LC-MS, GCxGC-MS, GC-MSxMS, GCxGC-MSxMS, LC-MSxMS, and NMR. By integrating data science techniques, the method offers a non-invasive, efficient, easy, and highly accurate approach for the early detection of diseases. This system is adaptable for a wide range of pathological states, leveraging biological samples such as urine, saliva, sebum, stool, sweat, and blood for disease prediction and screening.

[00083] Mass spectrometry assays, instruments and systems suitable for biomarker peptide analysis can include, without limitation, matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS; MALDI-TOF post-source-decay (PSD); MALDI-TOF/TOF; surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF) MS; electrospray ionization mass spectrometry (ESI-MS); ESI-MS/MS; ESI-MS/(MS)n (n is an integer greater than zero); ESI 3D or linear (2D) ion trap MS; ESI triple quadrupole MS; ESI quadrupole orthogonal TOF (Q-TOF); ESI Fourier transform MS systems; desorption/ionization on silicon (DIOS); secondary ion mass spectrometry (SIMS); atmospheric pressure chemical ionization mass spectrometry (APCI-MS); APCI-MS/MS; APCI-(MS)n; ion mobility spectrometry (IMS); inductively coupled plasma mass spectrometry (ICP-MS) atmospheric pressure photoionization mass spectrometry (APPI-MS); APPI-MS/MS; and APPI-(MS)n.

[00084] Mass spectrometers useful for the analyses described herein include, without limitation, Altis™ quadrupole, Quantis™ quadrupole, Quantiva™ or Fortis™ triple quadrupole from ThermoFisher Scientific, and the QSight™ Triple Quad LC/MS/MS from Perkin Elmer.

[00085] In various embodiments, mass spectrometry for analysis of VOCs can apply a wide detection range for diverse mass-to-charge ratios. In some embodiments, mass spectrometry for analysis of VOCs can apply a detection range of about 20 m/z to about 1000 m/z, about 20 m/z to about 950 m/z, about 20 m/z to about 900 m/z, about 20 m/z to about 850 m/z, about 20 m/z to about 800 m/z, about 20 m/z to about 750 m/z, about 20 m/z to about 700 m/z, about 20 m/z to about 650 m/z, about 20 m/z to about 600 m/z, about 20 m/z to about 550 m/z, about 20 m/z to about 500 m/z, about 20 m/z to about 450 m/z, or about 20 m/z to about 400 m/z. In some embodiments, mass spectrometry for analysis of VOCs can apply a detection range of about 50 m/z to about 550 m/z. In some embodiments, mass spectrometry for analysis of VOCs can apply a detection range of about 20 m/z to about 500 m/z.

III. Data and Analysis

[00086] (A) Measurements

[00087] Measurements produced by methods of the present disclosure that included chromatography followed by mass spectrometry can produce information including, for each measured analyte or VOC, mass-to-charge (m/z) ratio, retention time, and detection intensity. Such measurements can be analyzed by any of a variety of means known in the art. For example, data can be analyzed to determine the presence or absence of peaks of interest (optionally by application of a threshold), to determine the amounts or concentrations of molecules corresponding to one or more selected peaks (optionally where the amount or concentration is relative to an internal standard), or to provide a graph of all such peaks present in data set. Peaks can be, but are not necessarily, associated with a corresponding known or predicted molecule and/or VOC.

[00088] For example, in one embodiment, analytes, such as volatile organic compounds, can be fractionated in a first dimension by chromatography and, in a second dimension by mass spectrometry.

[00089] In one embodiment a dataset for analysis by the methods described herein comprises abundance data. Abundance data is the product of at least three dimensional analysis of volatiles in the sample. Volatile organic compounds in a sample are first fractionated in a first dimension by a chromatographic method, such as gas chromatography. Volatile organic compounds in different fractions coming off a chromatographic system are then further analyzed by mass spectrometry. The product of this analysis is a collection of mass spectra in which amounts of analytes are determined based on mass-to-charge ratio. Furthermore, to the extent mass spectrometry involves fragmenting volatile compounds, each peak on a mass spectrum can represent either a whole molecule or a fragment of a molecule.

[00090] When gas chromatography is coupled with mass spectrometry, mass spectrometer can sample fractions coming off the gas chromatograph and rates between about one per second and 100 per second, e.g., between about five per second and about 20 per second. Accordingly, if the gas chromatograph is run for 45 minutes, this could produce between about 2700 and about 270,000 mass spectra. Such process typically generates between about 1 million and 10 million data points. Each data point represents signal intensity at a particular mass to charge ratio at a particular retention time. (The combination of mass to charge ratio and retention time can be considered a feature of the overall data matrix.)

[00091] In addition, the dataset could include a chromatograph indicating measurements of analyte is a function of time.

[00092] In certain embodiments, the datasets comprise abundance data collected in three-dimensional or higher dimensional space. Such data can be generated by a sequence of fractionation of the sample and analysis of analytes in the fractions.

[00093] This disclosure contemplates creation of datasets with at least any of ten thousand, a hundred thousand, a million, or ten million or more data points.

[00094] A measurement of a variable can be any combination of numbers and words. A measure can be any scale, including nominal (e.g., name or category), ordinal (e.g., hierarchical order of categories), interval (distance between members of an order), ratio (interval compared to a meaningful "0"), or a cardinal number measurement that counts the number of things in a set. Measurements of a variable on a nominal scale indicate a name or category, e.g., category into which the sequencing read is classified. Measurements of a variable on an ordinal scale produce

a ranking, such as “first”, “second”, “third”. Measurements on a ratio scale include, for example, any measure on a pre-defined scale, absolute number of reads, normalized or estimated numbers, as well as statistical measurements such as frequency, mean, median, standard deviation, or quantile. Measurements that involve quantification are typically determined at the ratio scale level.

[00095] Data can be analyzed according to any of a variety of means, including by comparison to a reference standard generated by machine learning, a reference standard can be generated based on machine learning analysis of reference or training sample set representative of a disease or condition. In various embodiments, diagnosis, detection, and/or monitoring of a biological condition such as cancer can be achieved using a classification model.

[00096] **(B) Data Analysis and Classifying Categorical State**

[00097] In various embodiments, data generated according to a diagnostic method set forth herein (e.g., including GC-MS) can be analyzed according to any of a variety of methods. In some embodiments, one or more selected peaks are analyzed for their presence or absence, e.g., based on a threshold. In some embodiments, one or more selected peaks are analyzed based on their abundance (e.g., measured or relative amount or concentration, e.g., relative to an internal standard). In some embodiments, data are analyzed by unbiased comparison to a reference or standard, e.g., a reference or standard generated from a reference or training data set. In some embodiments, data are classified by comparison to a reference or standard generated from analysis of a reference or training data set using machine learning.

[00098] Certain embodiments of the compositions, methods, and systems can comprise: Providing a machine learning algorithm; acquiring biological samples, e.g., a biofluid as urine or sebum or spittle; tagging the samples with pathological states; performing metabolomic analysis by separating analytes in the samples and performing spectrographic analysis of the separated analytes, to produce an output; and training the algorithm with the output to associate patterns in the output with a pathological state (e.g., cancer).

[00099] Methods of using a classifier produced by the learning algorithm can involve: Analyzing biological samples using the analytical techniques described above to perform metabolomic analysis, and intelligently stopping run-time during analysis; and using the classifier to predict pathological states of the samples.

[000100] Using a classifier as described above, an operator can classify the categorical state of a particular categorical variable of a subject based on mass spectrometry data from the subject. The classifier can classify conditions according to any classification scheme useful to the operator. This can include, for example, a binary classification, such as present or absent or, into a plurality of categories, such as different subtypes of disease or different stages of disease.

[000101] (B)(1) Reference or Training Data Sets

[000102] Methods of generating models to predict a categorical state can involve providing a training dataset on which a machine learning algorithm can be trained to develop one or more models to predict or infer a categorical state. The training dataset will include a plurality of training examples, typically for each of a plurality of subjects and typically in the form of a vector. Each training example will include a plurality of features or data points and, for each feature, data, e.g., in the form of numbers or descriptors. Where learning is to be supervised, the data will include a classification of the sample or subject into a category of a categorical variable to be inferred. For example, the categorical variable may be "cancer diagnosis" and the categories or classifications of this variable can be "present" and "absent". Typically, for machine learning, the training examples will have at least 10, at least 100, at least 1000, at least 10,000, at least 100,000, at least 1,000,000, or at least 10,000,000 different features. The features selected are those on which prediction will be based.

[000103] (B)(2) Model Generation and Predicting Categorical State

[000104] In building or executing a model to predict the categorical state of a sample or an individual subject from whom the samples have been taken, datasets are provided that include information about one or a plurality of subjects.

[000105] Values for features in the dataset can be quantitative measures of the feature or descriptive terms. Quantitative measures can be given as a discrete or continuous range. Examples of quantitative measures include a number, a degree, a level, a range, percentage (*e.g.*, ratio, a numerator divided by a denominator, *etc.*) or bucket. A number can be a number on a scale, for example 1-10. Alternatively, the score can embrace a range. For example, ranges can be high, medium and low; severe, moderate and mild; or actionable and non-actionable. Buckets can comprise discrete numerals, such as 1-3, 4-6 and 7-10.

[000106] In GC-MS data, the measures typically will be intensity of signal a particular mass-to-charge ratio.

[000107] Models can be created by statistical methods, including, for example, methods performed by machine learning. Machine learning involves training machine learning algorithms on training data sets comprising data from a plurality of test subjects.

[000108] Methods for generating models to predict categorical state can comprise the following operations. A dataset as described above is provided. The dataset includes, for each of a plurality of subjects, abundance data as described herein. The data set is used as a training dataset to train a machine learning algorithm to produce one or more models that predict categorical state of a subject based on features in the abundance data. In some embodiments, the training dataset uses all of the abundance data or about all of the abundance data for training. In some embodiments, from an original dataset of between 1 million and 10 million feature data points per sample, a number of features representing for example, about two orders of magnitude fewer feature data points may be selected by the machine learning algorithm as being correlated with the categorical state. Among these features, another order of magnitude lower number may suffice in the final classification algorithm to produce a useful classifier.

[000109] Abundance data can be analyzed as raw data. In this case, the dataset is not reduced, for example, by extracting higher order information from the data, such as, particular molecular entities. Furthermore, analysis can proceed on all the raw data or on a portion of the raw data.

[000110] In certain embodiments of the methods disclosed herein the machine learning algorithm can develop the classifier that uses raw abundance data to identify features that correlate with a categorical state. In such case, the molecular identity of the feature or features that function as biomarkers may remain unknown.

[000111] In other embodiments, the raw data set is subject to data reduction. In this case, features or combinations of features are identified as corresponding to distinct molecular entities, e.g., VOCs or metabolites. These distinct molecular entities, rather than the raw feature data, can function as biomarkers in a classifier. Without wishing to be limited by theory, raw metabolism and data may function better in the development of classifiers because they can be normalized across many characteristics of subject, such as, for example, age, sex, race, geographic origin, or socioeconomic status.

[000112] (B)(3) Machine Learning Algorithms

[000113] The machine learning algorithm can be any suitable supervised machine learning algorithm, parametric or non-parametric. Machine learning algorithms include, without limitation, artificial neural networks (e.g., back propagation networks), decision trees (e.g., recursive partitioning processes, CART), random forests, discriminant analyses (e.g., Bayesian classifier or Fischer analysis), linear classifiers (e.g., multiple linear regression (MLR), partial least squares (PLS) regression, principal components regression (PCR)), mixed or random-effects models, non-parametric classifiers (e.g., k-nearest neighbors), support vector machines, and ensemble methods (e.g., bagging, boosting).

[000114] Using abundance data from a plurality of samples classified into the categorical states to be predicted, the machine learning algorithm generates a classifier that predicts, for an individual sample, the categorical state, based on abundance data collected from the sample.

[000115] In some embodiments, data in the dataset is not abstracted from the raw abundance data into higher order data, such as biomarkers. In the present case, a biomarker could be an individual, identified, volatile organic compound or specific fragment thereof. Accordingly, the abundance data represents a richer dataset for analysis then individual molecules themselves.

[000116] In certain embodiments, the machinery algorithm generates a confidence score in addition to the prediction. The confidence score indicates the confidence the classifier has that the prediction is correct. The confidence score threshold is a percentage value, a numerical value, a categorical value, ordinal value, a probability estimate (logistic regression, support vector machines, and neural networks), a softmax score (deep learning), a confidence interval, a margin (random forests), a prediction interval, an entropy-based measure, a Bayesian confidence measure (neural networks), a gini coefficient, an impurity measure (random forest), an anomaly score, or a ranking score. The confidence score can be, for example, an ordinal category such as “high”, “medium” and “low”.

[000117] Machine learning algorithms can be trained on the training dataset to generate models that predict the categorical state of a sample based on subject data, such as abundance data. Predicted categorical state can be translated into recommendations to a subject from home samples taken about, for example, health interventions, such as therapeutic treatments.

[000118] In various embodiments, a biological state is determined as a categorical variable. A categorical variable is a variable that can be defined by two or more, typically nonoverlapping, categorical states. In certain embodiments of the methods disclosed herein, the categorical variable is a pathological condition having two or more pathological states. For example, the pathological variable could be cancer and the categorical states could be positive or negative diagnosis. Alternatively, the categorical states could be stages of a disease, such as cancer, such as, stage 0, stage 1, stage 2, stage 3 or stage 4.

[000119] **(C) Dynamic Termination**

[000120] Gas chromatography on the sample can continue for, for example, about 45 minutes. However, it may not be necessary to collect data for 45 minutes on each subject sample in order to obtain a confident prediction of the categorical state. Provided herein are methods of more efficiently predicting a categorical state for sample from a subject using dynamic termination of data collection.

[000121] According to one method, a GC-MS run is performed on a sample from a subject. Spectra can be collected over a certain retention time and then the data collected can be used as a dataset for classification by classifier. The classifier outputs both a predicted categorical state and a confidence score for the predicted categorical state. If the confidence score is below a threshold, more spectra are collected from the sample over an extended retention time and the spectra are combined with the existing dataset. Then, the updated dataset is analyzed by the classifier to generate a new prediction and confidence score. If this confidence score is also, below the threshold, data collection continues to create another updated dataset which is, again, subject to classification and confidence prediction. When the confidence score meets or exceeds the threshold data collection can stop. The prediction and confidence score can be reported to the user. A New sample can now be loaded for analysis.

[000122] Accordingly, data is collected for the first period of time and used to generate a prediction and confidence score. If the confidence score is above a threshold data collection can stop. If the confidence score is below the threshold, data can continue to be collected over a second period of time to produce a large dataset. This dataset is also used to generate a prediction and confidence score. The process can continue until the confidence threshold is

achieved or, if it is predicted that the confidence score will not meet the threshold, for example, because improvements in the confidence score will plateau below the threshold.

[000123] The confidence score threshold can be selected by the operator. For example, the confidence score may be selected that indicates action should be taken. The confidence score may represent a bias in favor of reducing false positives or reducing false negatives.

[000124] In some cases, even after several iterations a confidence score meeting the threshold may not be reached. Therefore, the system can determine whether improvements in confidence score have plateaued. In this case, data collection can cease, and the resulting prediction and confidence score can be reported out. Improvements can be considered to have plateaued when the rate of improvement over several iterations approaches zero. A plateau is observed when improvements in the confidence score change over time drops below a certain threshold, or the forecasted confidence score over time for data to be collected is predicted below a certain threshold.

IV. Systems

[000125] Also provided herein are systems comprising a computer. Such systems can be used for, among other things, executing learning algorithms, executing classification algorithms to predict categorical state. Computer systems can include a central processing unit (also referred to as a CPU or a processor) memory (e.g., random-access memory, read-only memory, flash memory), communication interface for communicating with one or more other systems, and peripheral devices.

[000126] Such systems can be connected through a communications network to the Internet. The communications network can be any available network that connects to the Internet. The communication network can utilize, for example, a high-speed transmission network including, without limitation, Digital Subscriber Line (DSL), Cable Modem, Fiber, Wireless, Satellite and, Broadband over Powerlines (BPL).

[000127] Systems can include non-transitory computer readable medium that can contain machine-executable code that, upon execution by a computer processor, implements a method of the present disclosure.

V. Methods of Diagnosis and Treatment

[000128] The present disclosure includes the surprising discovery that VOCs in a sample from a subject may be detected, analyzed, and/or monitored according to methods disclosed herein to provide exceptional results for the detection, diagnosis, prediction of a diagnosis, and/or monitoring of a biological condition (e.g., cancer).

[000129] In certain embodiments, the categorical state is a pathological state or condition. Pathological conditions that can be categorized by the compositions, methods and systems of this disclosure include, without limitation, cancer, infectious disease, autoimmune disease, cardiovascular disease, metabolic disorder, neurological disorder, respiratory disease, gastrointestinal disease, musculoskeletal disorder, psychiatric/behavioral conditions, and dermatological conditions. Determining a categorical state of a pathological condition can also be referred to a "diagnosing" or "making a diagnosis."

[000130] Cancers that can be diagnosed include, without limitation, bladder cancer, breast cancer, cancer of the central nervous system, cervical cancer, colorectal cancer, endometrial cancer, esophageal cancer, glioblastoma, kidney cancer, leukemia, liver cancer (i.e., hepatocarcinoma), lung cancer (e.g., non-small cell lung cancer or NSCLC), melanoma, ovarian cancer, pancreatic cancer, prostate cancer, low-risk prostate cancer, high-risk prostate cancer, renal cancer (*e.g.*, renal cell carcinoma), skin cancer, stomach (gastric) cancer, testicular cancer, thyroid cancer, and uterine cancer.

[000131] In some embodiments, when the subject is diagnosed with a pathological condition by the methods described herein, the subject can be treated with a therapeutic intervention to improve health of the subject. Such therapeutic intervention can include administering to the subject a pharmaceutical composition specific for the disease state. For example, in the case of cancer, the therapeutic intervention could include administration of a chemotherapy drug, radiation, surgery, immunotherapy, stem cell transplant, hormone therapy or photodynamic therapy.

[000132] Classifications indicative of a biological condition can be provided to a subject, for example, in the form of recommendations. In one embodiment, the recommendations include a positive recommendation to administer a therapeutic intervention, such as a drug to treat the condition identified in a subject. In some embodiments, a subject is diagnosed as having a

cancer by a method provided herein, e.g. analysis of urine by GC-MS to detect VOCs, and subsequent to such diagnosis is treated by administration of a therapeutic agent for treatment of cancer (e.g., a chemotherapeutic agent). Accordingly, compositions, methods, and systems disclosed herein can be used in a method of treatment for a biological condition.

VI. Systems and Methods for generating a disease state prediction with a Machine-Learning Model

[000133] **Figure 6A** is a diagram **600** illustrating a system for generating disease prediction(s) **628** from mass spectrometry reads **602** according to embodiments of the present disclosure. The system can read an abundance matrix **608** that includes mass spectrometry reads **602** and mass spectrometry timestamp **604** dimensions. In some embodiments, the abundance matrix **608** comprises dimensions including a timestamp dimension, a mass-to-charge ratio dimension and an abundance dimension. In some embodiments, an optional abundance matrix generator **606** can combine tabular data, such as mass spectrometry reads **602** and mass spectrometry timestamp **604** into the abundance matrix **608**. In some embodiments, the abundance matrix **608** can be compiled by a mass spectrometry service, or a mass spectrometry data analytics service. In some embodiments, each observation in the abundance matrix comprises at least measurements (*e.g.*, dimensions) of retention time, m/z ratio, and intensity.

[000134] A trained ML model **626** can receive the abundance matrix **608.** The trained ML model **626** can be trained on a training set of abundance matrices and group truth disease states (*e.g.*, both healthy and with diseases) of subjects, where the abundance matrices were generated by a mass spectrometry analysis of samples of each of those subjects. In other words, each abundance matrix **Figures 7A** and **11A** illustrate a system and process for training such a model according to embodiments of the present disclosure in further detail. In relation to **Figure 6**, the trained ML **626** outputs predictions **628** based on the abundance matrix **608.** The predictions **628** can be predictions of disease state. In some embodiments, disease states can be the presence of conditions such as bladder cancer, kidney cancer, and prostate cancer. In some embodiments, disease states can be the presence of additional conditions such as colorectal cancer, breast cancer, or other cancers. In some embodiments, individual models are trained for each particular disease indication (*e.g.*, a first model for bladder cancer, a second model for kidney cancer, a

third model for prostate cancer, *etc.*). In some embodiments, a single model can be trained for multiple disease indications.

[000135] In some embodiments, the model can include one or more of a trained classifier, a trained regressor, a principal component analysis (PCA) model, a neural network, a support vector machine, random forest analysis, K-means clustering, Gaussian mixture models, gradient boosting machines, decision trees, logistic regression, naïve bayes, k-nearest neighbors, decision tree, and support vector machine, or any of these models in combination. In some embodiments, two or more of these models can be trained and each model can be weighted. In some embodiments, the classifier can include a random forest classifier. In some embodiments, the regressor can include a random forest regressor.

[000136] In some embodiments, an optional patient or doctor dashboard **630** can receive the prediction **628**. The dashboard **630** can display one or more of the predictions in a graphical or table format. In some embodiments, the dashboard **630** can display the diagnosis as a risk score. In some embodiments, the dashboard **630** can display a diagnoses, a quality control metric (*e.g.*, a fraction of biomarkers passed), a display of key biomarkers, a biomarker profile, and a model reasoning widget with a SHAP Waterfall Plot. In some embodiments, the SHAP Waterfall Plot displays a number of biomarkers and corresponding arrows, where a larger arrow indicates an importance of the biomarker. An example embodiment of the dashboard **630** can be illustrated in **Figures 8A, 8B, and 9** and is described in further detail below.

[000137] **Figure 6B** is a diagram **670** illustrating a system for generating disease prediction(s) **628** from mass spectrometry reads **602** according to embodiments of the present disclosure. The system can read an abundance matrix **608** that includes mass spectrometry reads **602** and mass spectrometry timestamp **604** dimensions. In some embodiments, the abundance matrix **608** comprises dimensions including a timestamp dimension, a mass-to-charge ratio dimension and an abundance dimension. In some embodiments, an optional abundance matrix generator **606** can combine tabular data, such as mass spectrometry reads **602** and mass spectrometry timestamp **604** into the abundance matrix **608**. In some embodiments, the abundance matrix **608** can be compiled by a mass spectrometry service, or a mass spectrometry data analytics service. In some embodiments, each observation in the abundance matrix comprises at least measurements (*e.g.*, dimensions) of retention time, m/z ratio, and intensity.

[000138] In some embodiments, an optional quality metric calculator **672** can analyze tabular data within the abundance matrix **608** and generate a quality score on a per sample basis and append those scores to a tabular file storing the abundance matrix **608**. In some embodiments, the quality metric calculator **672** can analyze the abundance matrix **608** and generate a quality score on a per sample basis and append those scores to the abundance matrix **608**.

[000139] A trained machine-learning (ML) model **626** can receive the tabular data of the abundance matrix **608** (*e.g.*, features representing biomarkers). The trained ML model **626** can be trained on a training set of tabular data from abundance matrices and group truth disease states (*e.g.*, both healthy and with diseases) of subjects, where the abundance matrices were generated by a mass spectrometry analysis of samples of each of those subjects. The tabular data can include **Figures 7B** and **11B** illustrate a system and process for training such a model according to embodiments of the present disclosure in further detail. In relation to **Figure 6B**, the trained ML **626** outputs predictions **628** based on the abundance matrix **608.** The predictions **628** can be predictions of disease state. In some embodiments, disease states can be the presence of conditions such as bladder cancer, kidney cancer, and prostate cancer. In some embodiments, disease states can be the presence of additional conditions such as colorectal cancer, breast cancer, or other cancers. In some embodiments, individual models are trained for each particular disease indication (*e.g.*, a first model for bladder cancer, a second model for kidney cancer, a third model for prostate cancer, *etc.*). In some embodiments, a single model can be trained for multiple disease indications.

[000140] In some embodiments, the model can include one or more of a trained classifier, a trained regressor, a principal component analysis (PCA) model, a neural network, a support vector machine, random forest analysis, K-means clustering, Gaussian mixture models, gradient boosting machines, decision trees, logistic regression, naïve bayes, k-nearest neighbors, decision tree, and support vector machine, or any of these models in combination. In some embodiments, two or more of these models can be trained and each model can be weighted. In some embodiments, the classifier can include a random forest classifier. In some embodiments, the regressor can include a random forest regressor.

[000141] In some embodiments, an optional patient or doctor dashboard **630** can receive the prediction **628**. The dashboard **630** can display one or more of the predictions in a graphical or

table format. In some embodiments, the dashboard **630** can display the diagnosis as a risk score. In some embodiments, the dashboard **630** can display a diagnoses, a quality control metric (*e.g.*, a fraction of biomarkers passed), a display of key biomarkers, a biomarker profile, and a model reasoning widget with a SHAP Waterfall Plot. In some embodiments, the SHAP Waterfall Plot displays a number of biomarkers and corresponding arrows, where a larger arrow indicates an importance of the biomarker. An example embodiment of the dashboard **630** can be illustrated in **Figures 8A, 8B, and 9** and is described in further detail below.

[000142] **Figure 6C** is a block diagram **680** illustrating a system for generating disease prediction(s) **628** from mass spectrometry reads **602** according to embodiments of the present disclosure. In some embodiments, an abundance matrix generator **606** loads mass spectrometry reads **602** and mass spectrometry timestamp data **604.** In some embodiments, the mass spectrometry timestamp data **604** can be either extracted from the mass spectrometry reads **602** or read separately from the mass spectrometry reads **602.** In some embodiments, the mass spectrometry reads **602** and the mass spectrometry timestamp data **604** are tensors, such as vectors. In some embodiments, the abundance matrix generator can convert the mass spectrometry reads **602** and the mass spectrometry timestamp data **604** to tensors, such as vectors. The abundance matrix generator **606** further generates an abundance matrix by combining the mass spectrometry reads **602** with the mass spectrometry timestamp data **604**, thereby generating the abundance matrix **608**. In some embodiments, each observation in the abundance matrix comprises at least measurements (*e.g.*, dimensions) of retention time, m/z ratio, and intensity.

[000143] A biomarker matching module **614** can load a plurality of biomarkers **612** from a database of biomarkers **610**. The biomarker matching module **614** can then output matched biomarkers **616**. In some embodiments, the matched biomarkers **616** can include biomarkers that have not been identified by a biomarker database **610**, but are identified via mass spectrometry data. A quality metric calculator **618** can then generate quality metrics **620** for each of the matched biomarkers **616**. A biomarker filter **622** can select filtered biomarkers **624** from the matched biomarkers **616**. The biomarker filter **622** can select filtered biomarkers **624** that have a quality metric over a particular threshold.

[000144] In some embodiments, the quality metric calculator **618** can calculate a quality metric that is based on one of the following: Mahalanobis Distance, Gaussian Mixture Model (GMM) Log Likelihood, Kolmogorov-Smirnov (KS) Test, Anderson-Darling Test, T-Test for Mean Differences, Principal Component Analysis (PCA) for Dimensionality Reduction & Visualization, Isolation Forest for Anomaly Detection, Calibration Curve & Logistic Regression, Dynamic Retention Time Termination in GC-MS, Peak Signal-to-Noise Ratio Assessment, Internal Standard Normalization, Feature Selection by Variance Thresholding, Outlier Detection by Robust Z-Score Analysis, Levene's Test for Homogeneity of Variance, Quality Metrics from Spectral Deconvolution (e.g., NIST Library Matching), Mass Spectrometry Data Clustering (K-Means, Hierarchical), Principal Component Regression (PCR), Random Forest Feature Selection, Gradient Boosting Model Evaluation, Variance Explained by Principal Components, SHAP Value Feature Importance for Biomarkers, Bayesian Inference for Outlier Detection, Machine Learning-Based Feature Selection, Peak Alignment and Retention Time Drift Correction, Metabolite Identification Confidence Scoring, Multi-Dimensional Scaling (MDS) for Batch Effects, Weighted Recursive Feature Elimination (WRFE), Euclidean Distance-Based Sample Similarity Scoring, Data Imputation Quality Metrics (kNN, Bayesian Methods), and Hierarchical Bayesian Models for Spectral Noise Estimation, Supervised vs. Unsupervised Feature Importance Comparison, Classifier Stability Across Bootstrapped Datasets, Mixed-Effects Modeling for Inter-Sample Variability, Spearman/Pearson Correlation Analysis of Replicates, Threshold-Based Peak Area Reproducibility, Dynamic Quality Control Filtering via Anomaly Detection, ROC Curve and AUC-Based Feature Evaluation, False Discovery Rate (FDR) Control in Biomarker Selection, Cross-Validation Performance Metrics (Accuracy, Precision, Recall, F1-score), Autoencoder-Based Feature Reduction & Anomaly Detection.

[000145] In some embodiments, the plurality of biomarkers **612** can be matched or determined based on one or more of the following data: (a) RT (min); (b) Scan number, Area (Ab*s); (c) baseline height (Ab); (d) Absolute Height; (e) Peak Width 50% (min); (f) Hit Number; (g) Hit Name; (h) Quality; (i) Mol Weight (amu); (j) CAS number; (k) Library file; (l) Entry number in the library. It can be recognized that biomarkers can also be identified without being in a library, for example.

[000146] A trained ML model **626** can receive the filtered biomarkers **624.** The trained ML model **626** can be trained on a training set of biomarkers and group truth disease states (*e.g.*, both healthy and with diseases) of subjects from which those biomarkers were sampled from. **Figure 7A** illustrates a system for training such a model according to embodiments of the present disclosure in further detail. In relation to **Figure 6**, the trained ML **626** outputs predictions **628** based on the set of filtered biomarkers **624.** The predictions **628** can be predictions of disease state. In some embodiments, disease states can be the presence of conditions such as bladder cancer, kidney cancer, and prostate cancer. In some embodiments, disease states can be the presence of additional conditions such as colorectal cancer, breast cancer, or other cancers. In some embodiments, individual models are trained for each particular disease indication (*e.g.*, a first model for bladder cancer, a second model for kidney cancer, a third model for prostate cancer, *etc.*). In some embodiments, a single model can be trained for multiple disease indications. In some embodiments, the disease state can be the presence of any cancer (*e.g.*, agnostic of the tissue origin) and the non-presence of any cancer (*e.g.*, agnostic of the tissue origin).

[000147] In some embodiments, the model can include one or more of a trained classifier, a trained regressor, a principal component analysis (PCA) model, a neural network, a support vector machine, random forest analysis, K-means clustering, Gaussian mixture models, gradient boosting machines, decision trees, logistic regression, naïve bayes, k-nearest neighbors, decision tree, and support vector machine, or any of these models in combination. In some embodiments, two or more of these models can be trained and each model can be weighted. In some embodiments, the classifier can include a random forest classifier. In some embodiments, the regressor can include a random forest regressor.

[000148] In some embodiments, an optional patient or doctor dashboard **630** can receive the prediction **628**. The dashboard **630** can display one or more of the predictions in a graphical or table format. In some embodiments, the dashboard **630** can display the diagnosis as a risk score. In some embodiments, the dashboard **630** can display a diagnoses, a quality control metric (*e.g.*, a fraction of biomarkers passed), a display of key biomarkers, a biomarker profile, and a model reasoning widget with a SHAP Waterfall Plot. In some embodiments, the SHAP Waterfall Plot displays a number of biomarkers and corresponding arrows, where a larger arrow indicates an

importance of the biomarker. An example embodiment of the dashboard **630** can be illustrated in **Figures 8A, 8B, and 9** and is described in further detail below.

[000149] **Figure 10A** is a flowchart **1000** illustrating a process of generating a disease prediction according to embodiments of the present disclosure. In some embodiments, the method includes loading an abundance matrix that represents mass spectrometry reads of a sample of a plurality of volatile organic compounds (VOCs) extracted from a biological sample (**1002**). The abundance matrix can represent each of the plurality of VOCs in a mass-to-charge ratio dimension, an abundance dimension, and a retention time dimension. The method can include providing the abundance matrix to a machine-learning model (**1004**). The machine-learning model can be trained on abundance matrixes of biomarkers of VOCs collected from healthy and diseased subjects. The method can include receiving, from the machine-learning model, a prediction of a state of a disease state (**1006**).

[000150] **Figure 10B** is a flowchart **1050** illustrating a process of generating a disease prediction according to embodiments of the present disclosure. The process can include loading tabular data of mass spectrometry reads that represent a sample of a plurality of volatile organic compounds (VOCs) extracted from a biological sample (**1052**). The mass spectrometry reads of the tabular data can represent each of the plurality of VOCs. The tabular data can comprise a plurality of tensors, the tensors including first tensor comprising a mass-to-charge ratio dimension, a second tensor comprising an abundance dimension, and a third tensor comprising a retention time dimension. The process can include extracting a first plurality of features (*e.g.*, features representing biomarkers) from the tabular data (**1054**). Each feature (*e.g.*, features representing biomarkers) can be associated with one of the VOCs. The process can include generating, for each of the first plurality of features (*e.g.*, features representing biomarkers), a quality metric based on the abundance matrix (**1056**). The process can include selecting a second plurality of features (*e.g.*, features representing biomarkers) from a set of the first plurality of features (*e.g.*, features representing biomarkers) that have a quality score above a threshold (**1058**). The process can include providing the second plurality of features (*e.g.*, features representing biomarkers) to a machine-learning model, the machine-learning model trained on features of biomarkers of VOCs collected from healthy and diseased subjects (**1060**).

The process can include receiving, from the machine-learning model, a prediction of a state of a disease state (**1062**).

VII. Systems and Methods for training a Machine-Learning Model to generate disease state predictions

[000151] **Figure 7A** is a block diagram **700** illustrating a system for training a machine learning model to generate disease prediction(s) from a training set of abundance matrixes **708.** Each abundance matrices includes mass spectrometry reads and is associated with subject disease data **726** according to embodiments of the present disclosure. A training set database **702** (*e.g.*, stored in a database) stores the abundance matrices **708** that store mass spectrometry reads and related data based on mass spectrometry performed on liquid biopsies from a plurality of subjects. The subject disease data **726** stored in the training set database **702** corresponds to each of those subjects from which the liquid biopsies are drawn. Each abundance matrix is originally generated from a mass spectrometry read of a sample (*e.g.*, a biological sample, a urine sample) received from a subject (*e.g.*, a patient, an individual). Each abundance matrix **708** is therefore associated with that subject. In some embodiments, each observation in the abundance matrix comprises at least measurements (*e.g.*, dimensions) of retention time, m/z ratio, and intensity.

[000152] In some embodiments, a model training factory **728** is configured to train a ML model **730**. The model training factory **728** can be a server including a processor and a memory, or a plurality of servers, each including a processor and a memory. The model training factory loads the abundance matrix **708** for each subject and also the subject disease data **726** for each subject. The model training factory **728** trains a ML model **730** based on this training set, thereby generating the ML model **730** that can receive an abundance matrix representing biomarkers and generate a disease state prediction.

[000153] In some embodiments, the ML model **730** can be an unsupervised model and trained with unsupervised machine learning methods. By clustering samples based on volatile organic compound (VOC) profiles, previously unrecognized pathological states or subtypes of diseases that were thought to be the same can be determined. Unsupervised methods in untargeted volatomics can uncover hidden patterns, enabling the discovery of new diseases. Once sufficient

data is gathered, unique clusters of biomarkers can be detected. These clusters of biomarkers can reveal a new pathological state that may have been previously unrecognized. For example, clustering may uncover new cancer subtypes based on distinct VOC signatures.

[000154] Unsupervised methods in untargeted volatomics can uncover hidden patterns, enabling a deeper understanding of existing diseases. In some embodiments, this approach can also reveal subtle separations within conditions. For example, the models can determine that diseases like bladder cancer can comprise molecular subtypes with unique metabolic pathways. For example, training data can be labeled with a disease type, as described above. Within that disease type, unsupervised machine learning techniques can be applied. Those unsupervised machine learning techniques can reveal either that there is one cluster of biomarkers (*e.g.*, VOCs), or two or more distinct clusters of biomarkers (*e.g.*, VOCs). When there is one cluster, it verifies that the disease has no subtypes. When two or more distinct clusters emerge, and one cluster is linked to a known patient subtype, the other cluster can represent a different, previously undefined subtype. In some embodiments, this can apply to other disease statuses or conditions such as dementia or fibromyalgia. These diagnoses are often treated as or referred to as catch-all diagnoses, however, they may each represent several distinct diseases. Employing the systems and methods disclosed herein in with at least partially unsupervised machine-learning methods can reveal whether there are distinct diseases within those diagnoses.

[000155] In some embodiments, biomarkers can be identified by an identifier such as a Chemical Abstracts Service (CAS) number. However, biomarkers that have not yet been identified or discovered may have no identifier. In some embodiments, biomarkers that have not been identified may have no cast number. Therefore, it can be advantageous to identify these biomarkers in a computational manner.

[000156] In some embodiments, unknown biomarkers can be tracked and/or identified with the following steps. The term unknown biomarker refers to a VOC produced by the human body that has not yet been identified or categorized. First, an anchor biomarker having an anchor peak can be selected. The anchor biomarker can be a VOC that is known to be produced by the human body. Unknown biomarkers therefore be expressed by a function relating the peak of the unknown biomarker to the anchor biomarker. In some embodiments, the function can determine a difference of the peak of the unknown biomarker to the anchor biomarker. Therefore, each

unknown biomarker can be expressed by a relative distance of its peak to the peak of the anchor biomarker.

[000157] In some embodiments, unknown biomarkers can be processed without identification of individual biomarkers. For example, pattern recognition can be applied to the abundance matrix. Each observation in the abundance matrix comprises at least measurements (*e.g.*, dimensions) of retention time, m/z ratio, and intensity. The pattern recognition can be applied to those measurements. The patterns can be correlated to known biomarkers (*e.g.*, compared to biomarkers from a database). The remaining patterns can then be classified as individual biomarkers or clusters of biomarkers. These biomarkers can then be correlated with disease states, even without prior identification or classification.

[000158] In some embodiments, self-supervised machine-learning methods or semi-self-supervised machine-learning methods can be employed. Self-supervised learning can extract VOC patterns from unlabeled data, thereby improving biomarker discovery, disease classification, and anomaly detection.

[000159] In some embodiments, self-supervised learning can comprise feature learning. Feature learning can train on large VOC datasets (*e.g.*, loaded from a database) to identify biologically relevant signatures without labels of those conditions. In some embodiments, the feature learning can identify features (*e.g.*, signatures) of conditions without using any labels, and compare those identified features to labels of the VOC dataset(s). This training can continue, for example, until the identified features are assessed as being within a loss function.

[000160] In some embodiments, the models can promote early detection. Such early detection can flag anomalies in VOC profiles that may indicate early stage disease. In some embodiments, the models can provide multi-disease screening. Multi-disease screening can enable cross-disease learning, making VOC-based diagnostics more adaptable and generalizable. Therefore, by integrating embodiments of either or both of the unsupervised and self-supervised approaches described above, VOC-based disease discovery, precision diagnostics, and real-time health monitoring can be provided in these manners.

[000161] In some embodiments, the model can include one or more of a classifier, a trained regressor, or a principal component analysis (PCA) model, or any of these models in combination. In some embodiments, two or more of these models can be trained and each model

can be weighted. In some embodiments, the classifier can include a random forest classifier. In some embodiments, the regressor can include a random forest regressor.

[000162] **Figure 7B** is a block diagram **770** illustrating a system for training a machine learning model to generate disease prediction(s) from a training set of abundance matrixes **708.** Each abundance matrices includes mass spectrometry reads and is associated with subject disease data **726** according to embodiments of the present disclosure. A training set database **702** (*e.g.*, stored in a database) stores the abundance matrices **708** that store mass spectrometry reads and related data based on mass spectrometry performed on liquid biopsies from a plurality of subjects. The subject disease data **726** stored in the training set database **702** corresponds to each of those subjects from which the liquid biopsies are drawn. Each abundance matrix is originally generated from a mass spectrometry read of a sample (*e.g.*, a biological sample, a urine sample) received from a subject (*e.g.*, a patient, an individual). Each abundance matrix **708** is therefore associated with that subject. In some embodiments, each observation in the abundance matrix comprises at least measurements (*e.g.*, dimensions) of retention time, m/z ratio, and intensity. The mass spectrometry tabular data **778** can include these measurements/dimensions.

[000163] In some embodiments, an optional tabular data extractor **772** can extract tabular data, the mass spectrometry tabular data **704** from the abundance matrix **708**. An optional sample quality metric calculator **774** can calculate a quality score **776** for each VOC of each sample in the tabular data, thereby providing per-sample quality scores **776**.

[000164] In some embodiments, a model training factory **728** is configured to train a ML model **730**. The model training factory **728** can be a server including a processor and a memory, or a plurality of servers, each including a processor and a memory. The model training factory loads mass spectrometry tabular data **778** and optionally the per-sample quality scores **776** for each subject and also the subject disease data **726** for each subject. The model training factory **728** trains a ML model **730** based on this training set, thereby generating the ML model **730** that can receive a set of biomarkers and generate a disease state prediction.

[000165] In some embodiments, the ML model **730** can be an unsupervised model and trained with unsupervised machine learning methods. By clustering samples based on volatile organic compound (VOC) profiles, previously unrecognized pathological states or subtypes of diseases

that were thought to be the same can be determined. Unsupervised methods in untargeted volatomics can uncover hidden patterns, enabling the discovery of new diseases. Once sufficient data is gathered, unique clusters of biomarkers can be detected. These cluster of biomarkers can reveal a new pathological state that may have been previously unrecognized. For example, clustering may uncover new cancer subtypes based on distinct VOC signatures.

[000166] Unsupervised methods in untargeted volatomics can uncover hidden patterns, enabling a deeper understanding of existing diseases. In some embodiments, this approach can also reveal subtle separations within conditions. For example, the models can determine that diseases like bladder cancer can comprise molecular subtypes with unique metabolic pathways. For example, training data can be labeled with a disease type, as described above. Within that disease type, unsupervised machine learning techniques can be applied. Those unsupervised machine learning techniques can reveal either that there is one cluster of biomarkers (*e.g.*, VOCs), or two or more distinct clusters of biomarkers (*e.g.*, VOCs). When there is one cluster, it verifies that the disease has no subtypes. When two or more distinct clusters emerge, and one cluster is linked to a known disease subtype, the other cluster can represent a known disease subtype or a different, previously undefined disease subtype. In some embodiments, this can apply to other disease statuses or conditions such as dementia or fibromyalgia. These diagnoses are often treated as or referred to as catch-all diagnoses, however, they may each represent several distinct diseases. Employing the systems and methods disclosed herein in with at least partially unsupervised machine-learning methods can reveal whether there are distinct diseases within those diagnoses.

[000167] In some embodiments, biomarkers can be identified by an identifier such as a Chemical Abstracts Services (CAS) number. However, biomarkers that have not yet been identified or discovered may have no identifier. In some embodiments, biomarkers that have not been identified may have no cast number. Therefore, it can be advantageous to identify these biomarkers in a computational manner.

[000168] In some embodiments, unknown biomarkers can be tracked and/or identified with the following steps. The term unknown biomarker refers to a VOC produced by the human body that has not yet been identified or categorized. First, an anchor biomarker having an anchor peak can be selected. The anchor biomarker can be a VOC that is known to be produced by the

human body. Unknown biomarkers therefore be expressed by a function relating the peak of the unknown biomarker to the anchor biomarker. In some embodiments, the function can determine a difference of the peak of the unknown biomarker to the anchor biomarker. Therefore, each unknown biomarker can be expressed by a relative distance of its peak to the peak of the anchor biomarker.

[000169] In some embodiments, unknown biomarkers can be processed without identification of individual biomarkers. For example, pattern recognition can be applied to the abundance matrix. Each observation in the abundance matrix comprises at least measurements (*e.g.*, dimensions) of retention time, m/z ratio, and intensity. The pattern recognition can be applied to those measurements. The patterns can be correlated to known biomarkers (*e.g.*, compared to biomarkers from a database). The remaining patterns can then be classified as individual biomarkers or clusters of biomarkers. These biomarkers can then be correlated with disease states, even without prior identification or classification.

[000170] In some embodiments, self-supervised machine-learning methods or semi-self-supervised machine-learning methods can be employed. Self-supervised learning can extract VOC patterns from unlabeled data, thereby improving biomarker discovery, disease classification, and anomaly detection.

[000171] In some embodiments, self-supervised learning can comprise feature learning. Feature learning can train on large VOC datasets (*e.g.*, loaded from a database) to identify biologically relevant signatures without labels of those conditions. In some embodiments, the feature learning can identify features (*e.g.*, signatures) of conditions without using any labels, and compare those identified features to labels of the VOC dataset(s). This training can continue, for example, until the identified features are assessed as being within a loss function.

[000172] In some embodiments, the models can promote early detection. Such early detection can flag anomalies in VOC profiles that may indicate early stage disease. In some embodiments, the models can provide multi-disease screening. Multi-disease screening can enable cross-disease learning, making VOC-based diagnostics more adaptable and generalizable. Therefore, by integrating embodiments of either or both of the unsupervised and self-supervised approaches described above, VOC-based disease discovery, precision diagnostics, and real-time health monitoring can be provided in these manners.

[000173] In some embodiments, the model can include one or more of a classifier, a trained regressor, or a principal component analysis (PCA) model, or any of these models in combination. In some embodiments, two or more of these models can be trained and each model can be weighted. In some embodiments, the classifier can include a random forest classifier. In some embodiments, the regressor can include a random forest regressor.

[000174] **Figure 7C** is a block diagram **780** illustrating a system for training a machine learning model to generate disease prediction(s) from a training set of mass spectrometry reads **704** and subject disease data **726** according to embodiments of the present disclosure. A training set database **702** (*e.g.*, stored in a database) stores a plurality of mass spectrometry reads **704** extracted from liquid biopsies from a plurality of subjects and subject disease data **726** that corresponds to each of those subjects. The pre-processes each mass spectrometry read **704** using the systems and steps illustrated within the per read processing region **740**. Therefore, each mass spectrometry read **704** is processed in this manner, either iteratively or in parallel.

[000175] An abundance matrix generator **706** loads mass spectrometry reads **704** and mass spectrometry timestamp data **705.** In some embodiments, the mass spectrometry timestamp data **705** can be either extracted from the mass spectrometry reads **704** or read separately from the mass spectrometry reads **704.** In some embodiments, the mass spectrometry reads **704** and the mass spectrometry timestamp data **705** are tensors, such as vectors. In some embodiments, the abundance matrix generator can convert the mass spectrometry reads **704** and the mass spectrometry timestamp data **705** to tensors, such as vectors. The abundance matrix generator **606** further generates an abundance matrix by combining the mass spectrometry reads **704** with the mass spectrometry timestamp data **705**, thereby generating the abundance matrix **708**.

[000176] A biomarker matching module **714** can load a plurality of biomarkers **712** from a database of biomarkers **710**. The biomarker matching module **714** can then output matched biomarkers **716**. In some embodiments, the matched biomarkers **716** can include biomarkers that have not been identified by a biomarker database **710**, but are identified via mass spectrometry data. A quality metric calculator **718** can then generate quality metrics **720** for each of the matched biomarkers **716**. A biomarker filter **722** can select filtered biomarkers **724** from the matched biomarkers **716**. The biomarker filter **722** can select filtered biomarkers **722** that have a quality metric over a particular threshold.

[000177] Each mass spectrometry read **704** is originally generated by a sample (*e.g.*, a biological sample, a urine sample) received from a subject (*e.g.*, a patient, an individual). Each spectrometry read **704** is therefore associated with that subject. The per-read processing **740** is performed for each mass spectrometry read **704** of the training set, thereby generating filtered biomarkers **724** for each mass spectrometry read **704** and therefor each subject.

[000178] In some embodiments, a model training factory **728** is configured to train a ML model **730**. The model training factory **728** can be a server including a processor and a memory, or a plurality of servers, each including a processor and a memory. The model training factory loads the filtered biomarkers **724** for each subject and also the subject disease data **726** for each subject. In some embodiments, the model training factory **728** or computing node can generate a training dataset including filtered biomarkers **724** that are associated with subject disease data **726**. Each filtered biomarker **724** can be associated with a subject from which a biopsy is collected, as described above. Likewise, the subject disease data **726** can be disease data associated with each of those subjects. Therefore, the model training factory **728**, or other computing node, can generate a training dataset comprising the filtered biomarkers **724** that is associated with subject disease data **726** associated with each biomarker. The model training factory **728** trains a ML model **730** based on this training set, thereby generating the ML model **730** that can receive a set of biomarkers and generate a disease state prediction.

[000179] In some embodiments, the ML model **730** can be an unsupervised model and trained with unsupervised machine learning methods. By clustering samples based on volatile organic compound (VOC) profiles, previously unrecognized pathological states or subtypes of diseases that were thought to be the same can be determined. Unsupervised methods in untargeted volatomics can uncover hidden patterns, enabling the discovery of new diseases. Once sufficient data is gathered, unique clusters of biomarkers can be detected. These clusters of biomarkers can reveal a new pathological state that may have been previously unrecognized. For example, clustering may uncover new cancer subtypes based on distinct VOC signatures.

[000180] Unsupervised methods in untargeted volatomics can uncover hidden patterns, enabling a deeper understanding of existing diseases. In some embodiments, this approach can also reveal subtle separations within conditions. For example, the models can determine that diseases like bladder cancer can comprise molecular subtypes with unique metabolic pathways.

For example, training data can be labeled with a disease type, as described above. Within that disease type, unsupervised machine learning techniques can be applied. Those unsupervised machine learning techniques can reveal either that there is one cluster of biomarkers (*e.g.*, VOCs), or two or more distinct clusters of biomarkers (*e.g.*, VOCs). When there is one cluster, it verifies that the disease has no subtypes. When two or more distinct clusters emerge, and one cluster is linked to a known patient subtype, the other cluster can represent a different, previously undefined subtype. In some embodiments, this can apply to other disease statuses or conditions such as dementia or fibromyalgia. These diagnoses are often treated as or referred to as catch-all diagnoses, however, they may each represent several distinct diseases. Employing the systems and methods disclosed herein in with at least partially unsupervised machine-learning methods can reveal whether there are distinct diseases within those diagnoses.

[000181] In some embodiments, biomarkers can be identified by an identifier such as a Chemical Abstracts Services (CAS) number. However, biomarkers that have not yet been identified or discovered may have no identifier. In some embodiments, biomarkers that have not been identified may have no cast number. Therefore, it can be advantageous to identify these biomarkers in a computational manner.

[000182] In some embodiments, unknown biomarkers can be tracked and/or identified with the following steps. The term unknown biomarker refers to a VOC produced by the human body that has not yet been identified or categorized. First, an anchor biomarker having an anchor peak can be selected. The anchor biomarker can be a VOC that is known to be produced by the human body. Unknown biomarkers therefore be expressed by a function relating the peak of the unknown biomarker to the anchor biomarker. In some embodiments, the function can determine a difference of the peak of the unknown biomarker to the anchor biomarker. Therefore, each unknown biomarker can be expressed by a relative distance of its peak to the peak of the anchor biomarker.

[000183] In some embodiments, unknown biomarkers can be processed without identification of individual biomarkers. For example, pattern recognition can be applied to the abundance matrix. Each observation in the abundance matrix comprises at least measurements (*e.g.*, dimensions) of retention time, m/z ratio, and intensity. The pattern recognition can be applied to those measurements. The patterns can be correlated to known biomarkers (*e.g.*, compared to

biomarkers from a database). The remaining patterns can then be classified as individual biomarkers or clusters of biomarkers. These biomarkers can then be correlated with disease states, even without prior identification or classification.

[000184] In some embodiments, self-supervised machine-learning methods or semi-self-supervised machine-learning methods can be employed. Self-supervised learning can extract VOC patterns from unlabeled data, thereby improving biomarker discovery, disease classification, and anomaly detection.

[000185] In some embodiments, self-supervised learning can comprise feature learning. Feature learning can train on large VOC datasets (*e.g.*, loaded from a database) to identify biologically relevant signatures without labels of those conditions. In some embodiments, the feature learning can identify features (*e.g.*, signatures) of conditions without using any labels, and compare those identified features to labels of the VOC dataset(s). This training can continue, for example, until the identified features are assessed as being within a loss function.

[000186] In some embodiments, the models can promote early detection. Such early detection can flag anomalies in VOC profiles that may indicate early stage disease. In some embodiments, the models can provide multi-disease screening. Multi-disease screening can enable cross-disease learning, making VOC-based diagnostics more adaptable and generalizable. Therefore, by integrating embodiments of either or both of the unsupervised and self-supervised approaches described above, VOC-based disease discovery, precision diagnostics, and real-time health monitoring can be provided in these manners.

[000187] In some embodiments, the model can include one or more of a classifier, a trained regressor, or a principal component analysis (PCA) model, or any of these models in combination. In some embodiments, two or more of these models can be trained and each model can be weighted. In some embodiments, the classifier can include a random forest classifier. In some embodiments, the regressor can include a random forest regressor.

[000188] **Figure 11A** is a flowchart **1100** illustrating a method of training a machine-learning model according to embodiments of the present disclosure. The method can comprise loading a plurality of disease states (**1102**). Each disease state can be associated with one of a plurality of subjects. The method can comprise loading a plurality of abundance matrices (**1104**). Each abundance matrix can represent mass spectrometry reads of a sample of a plurality of volatile

organic compounds (VOCs) extracted from a biological sample from one of the plurality of subjects. The mass spectrometry read can represent each of the plurality of VOCs. Each abundance matrix can represent each of the plurality of VOCs in a mass-to-charge ratio dimension, an abundance dimension, and a retention time dimension. The method can include training a machine-learning model on the plurality of abundance matrices and the plurality of disease states (**1106**). The resulting machine-learning model can be configured to receive a second abundance matrix and output a prediction of a disease state.

[000189] **Figure 11B** is a flowchart **1150** illustrating a method of training a machine-learning model according to embodiments of the present disclosure. The method can include loading a plurality of disease states (**1152**). Each disease state can be associated with one of a plurality of subjects. The method can include loading a plurality of tabular data of mass spectrometry reads (**1154**). Each tabular data can represent a sample of a plurality of volatile organic compounds (VOCs) extracted from a biological sample from one of the plurality of subjects. The tabular data can comprise a plurality of tensors. The tensors can include a first tensor comprising a mass-to-charge ratio dimension, a second tensor comprising an abundance dimension, and a third tensor comprising a retention time dimension. The method can include, for each of the plurality of tabular data, extracting a first plurality of biomarkers from the abundance matrix, each biomarker associated with one of the VOCs, generating, for each of the first plurality of biomarkers, a quality metric based on the abundance matrix, and selecting a second plurality of biomarkers from a set of the first plurality of biomarkers that have a quality score above a threshold (**1156**). The method can include generating a training dataset comprising, for each subject, (a) the disease state associated with that subject and (b) the second plurality of biomarkers (**1158**). The method can include training a machine-learning model on the training dataset, the resulting machine-learning model configured to receive a third plurality of biomarkers and output a prediction of a disease state (**1160**).

[000190] In some embodiments, the tensors are loaded from a structured form. The data can be processed by matching peaks to a database (*e.g.*, the National Institute of Standards and Technology (NIST) database) and assigning predicted biomarkers and quality scores.

[000191] **Figures 8A** and **8B** are user interfaces **800** and **850** illustrating example outputs of a dashboard according to embodiments of the present disclosure. In **Figure 8A**, a history of

uploads illustrates a plurality of file uploads for different subjects (*e.g.*, patients, individuals) according to embodiments of the present disclosure. Each upload is associated in the dashboard with a prediction of a disease state for that subject's sample. For example, the first upload is for a 68 year old male and indicates positive for bladder cancer, but negative for kidney cancer and prostate cancer. As another example, the second upload is for a 76 year old male and indicates negative for bladder cancer, kidney cancer, and prostate cancer. As another example, the third upload is for a 68 year old male and indicates positive for bladder cancer, but negative for kidney cancer and prostate cancer. As another example, the fourth upload is for a 76 year old male and indicates negative for bladder cancer, kidney cancer, and prostate cancer. As another example, the fifth upload is for a 68 year old male and indicates positive for bladder cancer, but negative for kidney cancer and prostate cancer.

[000192] In **Figure 8B**, a history of uploads illustrates a plurality of file uploads for different subjects (*e.g.*, patients, individuals) according to embodiments of the present disclosure. In **Figure 8B**, four uploads and relative diagnoses are displayed. In addition to a positive and negative diagnoses, each positive diagnoses is rated with a ranked likelihood. The four uploads illustrated are exemplary, and other uploads or samples can be analyzed and displayed in the dashboard. The exemplary uploads include Patient 2, which shows negative for bladder cancer, kidney cancer, and prostate cancer; Patient 3, which includes a positive indication for bladder cancer, a positive indication for kidney cancer, and a negative indication for prostate cancer; Patient 4, which includes a positive indication for bladder cancer, kidney cancer, and prostate cancer, and Patent 5, which includes a positive indication for bladder cancer and prostate cancer and a negative indication for kidney cancer. In this example, Patients 3-5 include multiple positive indications for cancer, and therefore they also include a ranking of likelihood. Patient 3 includes an indication that bladder cancer is most likely (1^{st}) and kidney cancer is next to most likely (2^{nd}). Patient 4 includes an indication that kidney cancer is most likely (1^{st}), bladder cancer is next most likely (2^{nd}), and prostate cancer is third to most likely (3^{rd}). Patient 5 includes an indication that bladder cancer is most likely (1^{st}) and prostate cancer is next to most likely (2^{nd}).

[000193] In some embodiments, the rankings can be determined by a vote count method. Multiple models can be trained and used in a similar manner described above, but an be trained

using different settings, such as parameters or hyperparameters, or on different training datasets. Each model can provide a prediction of a disease state (including no disease state), and those predictions can be counted, where each model can provide one vote. In some embodiments, each model can provide a prediction of a disease state (including no disease state), and those predictions can be counted, where each model can provide a weighted vote. The sum of those votes for each condition for a given subject provides the ranked order shown in **Figure 8B**, where the most likely condition has the most votes, the second most likely has the second most votes, *etc*.

[000194] **Figure 9** is a user interface **900** illustrating a dashboard according to embodiments of the present disclosure. The dashboard can display two disease states detected for a subject's sample – prostate cancer and bladder cancer – and a plurality of widgets for each diagnosis. A first widget can illustrate disease state for each condition, which can be configured to display positive or negative, and optionally a ranking. A risk score widget can illustrate a certainty of the diagnosis. A quality control widget can indicate a number of biomarkers passed. A key biomarkers widget can illustrate one or more biomarkers that were influential in the prediction of the disease state in a SHAP Force Plot. A biomarker profile widget can display a violin plot illustrating biomarkers that were influential in the prediction of the disease state. A modeling reasoning widget can display a SHAP Waterfall Plot that illustrates the one or more model's reasoning.

VIII. Biomarker Selection for Model Training and Use

[000195] In some embodiments, the above systems and methods can be configured to use any biomarker present in the training datasets and samples. However, in some examples, all of these biomarkers may not be necessary for the models to give an accurate prediction. In some embodiments, training or using a model using fewer biomarkers can provide advantages such as using less computing power, higher speed of the model and/or higher speed of training the model. A validation process can ensure that the resulting model provides a statistically similar accuracy a model with more biomarkers included.

[000196] In some embodiments, the biomarker selection process can include processes including Exhaustive Enumeration of Combinations, Minimum Feature Frontier – Averages, a Minimum Feature Frontier – Blinding, or a combination of any of the above.

[000197] In some embodiments, an Exhaustive Enumeration of Combinations process can include running all feature combinations that surpass a threshold (*e.g.*, an AUC greater than or above 0.80).

[000198] In some embodiments, a Minimum Feature Frontier – Averages process can begin by ranking features (e.g., biomarkers) by importance (*e.g.*, VIP scores). Of the ranked features, a top *N* features can be selected as the pool from which to build models. Models can be constructed having feature subsets sized from *B* to *N*, build and test models, where *B* is a changing variable. Each of these models can be evaluated to determine how many exceed the AUC threshold (e.g., 0.80), and for which threshold *B* the models exceed the AUC threshold. Then, the process can determine where the number of features (1 to *N*) intersects (*Z*) with the AUC threshold (AUC = 0.80), where *Z* or more features are more likely to be provide an effective model, and fewer than *Z* features are less likely to provide an effective model. For example, when *N* is 10 features, the intersection may occur at 4 features. This may mean that any combination of 4 or more features (from the list of the top *N* features) with an AUC $\geq$ 0.80 will be likely to train an effective model. The intersection Z can be 10, 60% in this example.

[000199] In some embodiments, the importance metric can comprise one, or any combination, of the following scores and/or metrics: Variable Importance in Projection (VIP) Scores, SHAP (SHapley Additive Explanations) Values, Permutation Importance, Gini Importance (Mean Decrease in Impurity), Gain Importance (XGBoost), Weight Importance (LightGBM), CatBoost Feature Importance, LASSO (L1 Regularization Weights), Ridge Regression (L2 Regularization Weights), Elastic Net (L1 + L2 Regularization), Logistic Regression Coefficients, T-Test, ANOVA F-Test, Chi-Square Test, Mutual Information Score, Pearson Correlation Coefficient, Spearman Rank Correlation, Kendall Rank Correlation, Information Gain (Entropy Reduction), Kolmogorov-Smirnov (KS) Statistic, Decision Tree-Based Feature Ranking (Classification and Regression Trees (CART), Random Forest, XGBoost), Support Vector Machine (SVM) Feature Weights, Neural Network Feature Weights, Autoencoder-Based Feature Selection, Recursive Feature Elimination (RFE), Forward Feature Selection, Backward Feature Elimination, Stepwise

Feature Selection, Principal Component Analysis (PCA) Loadings, Independent Component Analysis (ICA) Weights, t-SNE Feature Contributions, UMAP Feature Contributions, Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), Bayesian Feature Ranking (Bayesian Networks), Expert-Driven Feature Ranking, Multi-Criteria Decision Analysis (MCDA), Feature Interaction Scores, ReliefF Algorithm, Fisher Score, Stability Selection, Wrapper Method Feature Selection, Boruta Algorithm, Minimum Redundancy Maximum Relevance (mRMR), Conditional Mutual Information Maximization (CMIM), F-Score Ranking, Greedy Feature Selection, Genetic Algorithm-Based Feature Ranking, Mutual Information Maximization, Recursive Feature Addition, Leave-One-Out Feature Ranking, Jackknife Resampling Importance, Random Subspace Feature Selection, Ensemble Feature Importance Averaging, Heterogeneity-Based Feature Importance, Fisher Discriminant Ratio, Cohen's D Effect Size Ranking, Mann-Whitney U Test Feature Ranking, Welch's T-Test Ranking, Mahalanobis Distance-Based Ranking, Quadratic Discriminant Analysis (QDA) Importance, and Canonical Corre.

[000200] In some embodiments, the process can iterate for different values of *N*. Repeat the above process for multiple values of N (1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, *etc.*). As *N* increases, the intersection threshold *Z* should shift upwards. For example, consider an example where *N* is 100 features. The intersection *Z* may increase because each feature contributes less to the model. In an example, if *Z* intersects at 80 features, the intersection can be represented as 100, 20%.

[000201] In some embodiments, an efficient frontier process is employed. In the efficient frontier process, the number of biomarkers in the model and the percentage of models that surpass the double-threshold (*e.g.*, 50% of models above 0.90 AUC) are considered. The number of biomarkers can be represented graphically on an X-axis and the percentage of models that surpass the double-threshold can be represented on the Y-axis. The curve (or "frontier") can illustrate how many subsets are above the threshold for each feature subset size.

[000202] In some embodiments, a Minimum Feature Frontier – Blinding process is employed. The top biomarkers can be blinded to find a threshold of accuracy. First, the process ranks biomarkers by establishing a list *L* of top biomarkers by importance (*e.g.*, with the VIP scores or other importance metrics listed herein).

[000203] In one example blinding iteration, a first iteration may blind 0 biomarkers. In this iteration, all biomarkers in L are used. This iteration finds the minimum number of features (F) needed to surpass the threshold (AUC $\geq$ 0.80 or 0.90, depending on the goal). A second iteration can then blind 1 biomarker. A top ranked biomarker is removed from L. Then, the process determines whether the resulting models still exceed the threshold with F features. If not, the method tries $F+1$, $F+2$, until a new number of features F is established as the frontier for this number of blinded biomarkers. Then, this process is repeated for each number of blinded biomarkers until the threshold cannot be passed.

[000204] In some embodiments, an efficient frontier (blinding) can be represented graphically. The X-axis can represent a number of biomarkers blinded (from 0 to the total count). The Y-axis can represent the minimum number of features from the remaining set needed to surpass the threshold. The curve starts slowly at the bottom-left, then rises to the right with accelerating growth.

[000205] In some embodiments, the AUC for any of the above methods can comprise 0.7, 0.8, and 0.85. In some embodiments, the AUC for any of the above methods can be at least any of the above: 0.75, 0.80, 0.81, 0.82, 0.83, 0.84, 0.85, 0.86, 0.87, 0.88, 0.89, 0.90, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.985, 0.99, 0.991, 0.992, 0.993, 0.994, 0.995, 0.996, 0.997, 0.998, and 0.999.

IX. An Exemplary Computing Node

[000206] Referring now to **Figure 13**, a schematic of an example of a computing node is shown. Computing node **10** is only one example of a suitable computing node and is not intended to suggest any limitation as to the scope of use or functionality of embodiments described herein. Regardless, computing node **10** is capable of being implemented and/or performing any of the functionality set forth hereinabove.

[000207] In computing node **10** there is a computer system/server **12**, which is operational with numerous other general purpose or special purpose computing system environments or configurations. Examples of well-known computing systems, environments, and/or configurations that may be suitable for use with computer system/server **12** include, but are not limited to, personal computer systems, server computer systems, thin clients, thick clients,

handheld or laptop devices, multiprocessor systems, microprocessor-based systems, set top boxes, programmable consumer electronics, network PCs, minicomputer systems, mainframe computer systems, and distributed cloud computing environments that include any of the above systems or devices, and the like.

[000208] Computer system/server **12** may be described in the general context of computer system-executable instructions, such as program modules, being executed by a computer system. Generally, program modules may include routines, programs, objects, components, logic, data structures, and so on that perform particular tasks or implement particular abstract data types. Computer system/server **12** may be practiced in distributed cloud computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed cloud computing environment, program modules may be located in both local and remote computer system storage media including memory storage devices.

[000209] As shown in **Figure 13**, computer system/server **12** in computing node **10** is shown in the form of a general-purpose computing device. The components of computer system/server **12** may include, but are not limited to, one or more processors or processing units **16**, a system memory **28**, and a bus **18** that couples various system components including system memory **28** to processor **16**.

[000210] Bus **18** represents one or more of any of several types of bus structures, including a memory bus or memory controller, a peripheral bus, an accelerated graphics port, and a processor or local bus using any of a variety of bus architectures. By way of example, and not limitation, such architectures include Industry Standard Architecture (ISA) bus, Micro Channel Architecture (MCA) bus, Enhanced ISA (EISA) bus, Video Electronics Standards Association (VESA) local bus, Peripheral Component Interconnect (PCI) bus, Peripheral Component Interconnect Express (PCIe), and Advanced Microcontroller Bus Architecture (AMBA).

[000211] Computer system/server **12** typically includes a variety of computer system readable media. Such media may be any available media that is accessible by computer system/server **12**, and it includes both volatile and non-volatile media, removable and non-removable media.

[000212] System memory **28** can include computer system readable media in the form of volatile memory, such as random access memory (RAM) **30** and/or cache memory **32**.

Computer system/server **12** may further include other removable/non-removable, volatile/non-volatile computer system storage media. By way of example only, storage system **34** can be provided for reading from and writing to a non-removable, non-volatile magnetic media (not shown and typically called a "hard drive"). Although not shown, a magnetic disk drive for reading from and writing to a removable, non-volatile magnetic disk (*e.g.*, a "floppy disk"), and an optical disk drive for reading from or writing to a removable, non-volatile optical disk such as a CD-ROM, DVD-ROM or other optical media can be provided. In such instances, each can be connected to bus **18** by one or more data media interfaces. As will be further depicted and described below, memory **28** may include at least one program product having a set (*e.g.*, at least one) of program modules that are configured to carry out the functions of embodiments of the disclosure.

[000213] Program/utility **40**, having a set (at least one) of program modules **42**, may be stored in memory **28** by way of example, and not limitation, as well as an operating system, one or more application programs, other program modules, and program data. Each of the operating system, one or more application programs, other program modules, and program data or some combination thereof, may include an implementation of a networking environment. Program modules **42** generally carry out the functions and/or methodologies of embodiments as described herein.

[000214] Computer system/server **12** may also communicate with one or more external devices **14** such as a keyboard, a pointing device, a display **24**, etc.; one or more devices that enable a user to interact with computer system/server **12**; and/or any devices (*e.g.*, network card, modem, etc.) that enable computer system/server **12** to communicate with one or more other computing devices. Such communication can occur via Input/Output (I/O) interfaces **22**. Still yet, computer system/server **12** can communicate with one or more networks such as a local area network (LAN), a general wide area network (WAN), and/or a public network (*e.g.*, the Internet) via network adapter **20**. As depicted, network adapter **20** communicates with the other components of computer system/server **12** via bus **18**. It should be understood that although not shown, other hardware and/or software components could be used in conjunction with computer system/server **12**. Examples, include, but are not limited to: microcode, device drivers,

redundant processing units, external disk drive arrays, RAID systems, tape drives, and data archival storage systems, etc.

[000215] The present disclosure may be embodied as a system, a method, and/or a computer program product. The computer program product may include a computer readable storage medium (or media) having computer readable program instructions thereon for causing a processor to carry out aspects of the present disclosure.

[000216] The computer readable storage medium can be a tangible device that can retain and store instructions for use by an instruction execution device. The computer readable storage medium may be, for example, but is not limited to, an electronic storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing. A non-exhaustive list of more specific examples of the computer readable storage medium includes the following: a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a static random access memory (SRAM), a portable compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk, a mechanically encoded device such as punch-cards or raised structures in a groove having instructions recorded thereon, and any suitable combination of the foregoing. A computer readable storage medium, as used herein, is not to be construed as being transitory signals per se, such as radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating through a waveguide or other transmission media (*e.g.*, light pulses passing through a fiber-optic cable), or electrical signals transmitted through a wire.

[000217] Computer readable program instructions described herein can be downloaded to respective computing/processing devices from a computer readable storage medium or to an external computer or external storage device via a network, for example, the Internet, a local area network, a wide area network and/or a wireless network. The network may comprise copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the

network and forwards the computer readable program instructions for storage in a computer readable storage medium within the respective computing/processing device.

[000218] Computer readable program instructions for carrying out operations of the present disclosure may be assembler instructions, instruction-set-architecture (ISA) instructions, machine instructions, machine dependent instructions, microcode, firmware instructions, state-setting data, or either source code or object code written in any combination of one or more programming languages, including an object oriented programming language such as Smalltalk, C++ or the like, and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The computer readable program instructions may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider). In some embodiments, electronic circuitry including, for example, programmable logic circuitry, field-programmable gate arrays (FPGA), or programmable logic arrays (PLA) may execute the computer readable program instructions by utilizing state information of the computer readable program instructions to personalize the electronic circuitry, in order to perform aspects of the present disclosure.

[000219] Aspects of the present disclosure are described herein with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems), and computer program products according to embodiments of the disclosure. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer readable program instructions.

[000220] These computer readable program instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks. These computer

readable program instructions may also be stored in a computer readable storage medium that can direct a computer, a programmable data processing apparatus, and/or other devices to function in a particular manner, such that the computer readable storage medium having instructions stored therein comprises an article of manufacture including instructions which implement aspects of the function/act specified in the flowchart and/or block diagram block or blocks.

[000221] The computer readable program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other device to cause a series of operational steps to be performed on the computer, other programmable apparatus or other device to produce a computer implemented process, such that the instructions which execute on the computer, other programmable apparatus, or other device implement the functions/acts specified in the flowchart and/or block diagram block or blocks.

[000222] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods, and computer program products according to various embodiments of the present disclosure. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of instructions, which comprises one or more executable instructions for implementing the specified logical function(s). In some alternative implementations, the functions noted in the block may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts or carry out combinations of special purpose hardware and computer instructions.

[000223] The descriptions of the various embodiments of the present disclosure have been presented for purposes of illustration, but are not intended to be exhaustive or limited to the embodiments disclosed. Many modifications and variations will be apparent to those of ordinary skill in the art without departing from the scope and spirit of the described embodiments. The terminology used herein was chosen to best explain the principles of the embodiments, the

practical application or technical improvement over technologies found in the marketplace, or to enable others of ordinary skill in the art to understand the embodiments disclosed herein.

EXEMPLARY EMBODIMENTS

[000224] The present disclosure provides, among other things, the following exemplary embodiments:

[000225] Embodiment 1. A method comprising:

loading an abundance matrix that represents mass spectrometry reads of a sample of a plurality of volatile organic compounds (VOCs) extracted from a biological sample, the abundance matrix representing each of the plurality of VOCs in a mass-to-charge ratio dimension, an abundance dimension, and a retention time dimension;

providing the abundance matrix to a machine-learning model, the machine-learning model trained on abundance matrixes of biomarkers of VOCs collected from healthy and diseased subjects; and

receiving, from the machine-learning model, a prediction of a state of a disease state.

[000226] Embodiment 2. The method of Embodiment 1, wherein the plurality of VOCs include at least one of VOCs having unidentified biomarkers.

[000227] Embodiment 3. The method of any of Embodiments 1-2, wherein the biological sample is a urine sample.

[000228] Embodiment 4. The method of any of Embodiments 1-3, wherein the disease state is one or more pathological state.

[000229] Embodiment 5. The method of any of Embodiments 1-4, wherein the disease state is one or more of kidney cancer, bladder cancers, and prostate cancer.

[000230] Embodiment 6. The method of any of Embodiments 1-5, further comprising:

displaying, a user interface, at least one widget of a dashboard, the widgets display at least one of a diagnosis, a quality control metric, biomarkers ranked above a threshold, a biomarker profile, and a model reasoning widget with a SHAP Waterfall Plot.

[000231] Embodiment 7. The method of any of Embodiments 1-6, further comprising:

generating the abundance matrix based on raw data received from a mass spectrometer.

[000232] Embodiment 8. A method comprising:

receiving, from each of a plurality of machine-learning models, a prediction of a state of a disease state according to the method of Embodiment 1, thereby resulting in a plurality of predictions, the plurality of predictions including at least a first disease state and a second disease state; and

ranking, based on the plurality of predictions, the first disease state and the second disease state.

[000233] Embodiment 9. The method of Embodiment 8, wherein ranking the first disease state and the second disease state includes summing the predictions of each disease state, and ranking the disease states by their summed total.

[000234] Embodiment 10. The method of Embodiment 8, wherein ranking the first disease state and the second disease state includes:

weighting the predictions of each disease state;

summing the weighted predictions of each disease state; and

ranking the disease states by their summed total.

[000235] Embodiment 11. A method comprising:

loading a plurality of disease states, each disease state associated with one of a plurality of subjects;

loading a plurality of abundance matrices, each abundance matrix representing mass spectrometry reads of a sample of a plurality of volatile organic compounds (VOCs) extracted from a biological sample from one of the plurality of subjects, the mass spectrometry read representing each of the plurality of VOCs, each abundance matrix representing each of the plurality of VOCs in a mass-to-charge ratio dimension, an abundance dimension, and a retention time dimension; and

training a machine-learning model on the plurality of abundance matrices and the plurality of disease states, the resulting machine-learning model configured to receive a second abundance matrix and output a prediction of a disease state.

[000236] Embodiment 12. The method of Embodiment 11, wherein the plurality of VOCs include at least one of VOCs having unidentified biomarkers.

[000237] Embodiment 13. The method of any of Embodiments 11-12, wherein the biological sample is a urine sample.

[000238] Embodiment 14. The method of any of Embodiments 11-13, wherein the disease state is a pathological state.

[000239] Embodiment 15. The method of any of Embodiments 11-14 wherein the disease state is one or more of kidney cancer, bladder cancer, and prostate cancer.

[000240] Embodiment 16. The method of any of Embodiments 11-15, further comprising:

generating the abundance matrix based on raw data received from a mass spectrometer.

[000241] Embodiment 17. A method comprising:

generating a first ranked list of biomarkers, the biomarkers ranked by an importance metric;

selecting a plurality of biomarkers from the ranked list, the plurality of biomarkers being the highest ranked biomarkers of the ranked list;

generating a second ranked list of biomarkers, the second ranked list comprising the biomarkers of the first ranked list with the plurality of biomarkers removed; and

training a machine learning model using the method of any of Embodiments 11-16, wherein the second plurality of biomarkers in the training dataset is the second ranked list of biomarkers.

[000242] Embodiment 18. A method comprising:

generating a ranked list of biomarkers, the biomarkers ranked by an importance metric;

selecting a plurality of biomarkers from the ranked list, the plurality of biomarkers being the highest ranked biomarkers of the ranked list; and

training a machine learning model using the method of any of Embodiments 11-16, wherein the second plurality of biomarkers in the training dataset is the plurality of biomarkers.

[000243] Embodiment 19. A method comprising:

generating a first ranked list of biomarkers, the biomarkers ranked by an importance metric;

selecting a plurality of biomarkers from the ranked list, the plurality of biomarkers being the highest ranked biomarkers of the ranked list;

generating a second ranked list of biomarkers, the second ranked list comprising the biomarkers of the first ranked list with the plurality of biomarkers removed;

generating a third ranked list of biomarkers, the third ranked list comprising a plurality of top ranked biomarkers from the second ranked list; and

training a machine learning model using the method of any of Embodiments 11-16, wherein the second plurality of biomarkers in the training dataset is the third ranked list of biomarkers.

[000244] Embodiment 20. A method comprising:

loading tabular data of mass spectrometry reads that represent a sample of a plurality of volatile organic compounds (VOCs) extracted from a biological sample, the mass spectrometry read representing each of the plurality of VOCs, the tabular data comprising a plurality of tensors, the tensors including first tensor comprising a mass-to-charge ratio dimension, a second tensor comprising an abundance dimension, and a third tensor comprising a retention time dimension;

extracting a first plurality of features from the tabular data, each feature of the first plurality associated with one of the VOCs;

generating, for each of the first plurality of features, a quality metric based on the abundance matrix;

selecting a second plurality of features from a set of the first plurality of features that have a quality score above a threshold;

providing the second plurality of features to a machine-learning model, the machine-learning model trained on features of VOCs collected from healthy and diseased subjects; and

receiving, from the machine-learning model, a prediction of a state of a disease state.

[000245] Embodiment 21. The method of Embodiment 20, wherein extracting the first plurality of features further includes comparing the tabular data to a database of features, the comparison resulting in the first plurality of features.

[000246] Embodiment 22. The method of any of Embodiments 20-21, wherein the plurality of VOCs include at least one of VOCs having unidentified biomarkers.

[000247] Embodiment 23. The method of any of Embodiments 20-22, wherein the biological sample is a urine sample.

[000248] Embodiment 24. The method of any of Embodiments 20-23, wherein the quality metric is calculated based on one or more of the following methods: Mahalanobis Distance, Gaussian Mixture Model (GMM) Log Likelihood, Kolmogorov-Smirnov (KS) Test, Anderson-Darling Test, T-Test for Mean Differences, Principal Component Analysis (PCA) for Dimensionality Reduction & Visualization, Isolation Forest for Anomaly Detection, Calibration Curve & Logistic Regression, Dynamic Retention Time Termination in GC-MS, Peak Signal-to-Noise Ratio Assessment, Internal Standard Normalization, Feature Selection by Variance Thresholding, Outlier Detection by Robust Z-Score Analysis, Levene's Test for Homogeneity of Variance, Quality Metrics from Spectral Deconvolution (e.g., NIST Library Matching), Mass Spectrometry Data Clustering (K-Means, Hierarchical), Principal Component Regression (PCR), Random Forest Feature Selection, Gradient Boosting Model Evaluation, Variance Explained by Principal Components, SHAP Value Feature Importance for Biomarkers, Bayesian Inference for Outlier Detection, Machine Learning-Based Feature Selection, Peak Alignment and Retention Time Drift Correction, Metabolite Identification Confidence Scoring, Multi-Dimensional Scaling (MDS) for Batch Effects, Weighted Recursive Feature Elimination (WRFE),

Euclidean Distance-Based Sample Similarity Scoring, Data Imputation Quality Metrics (kNN, Bayesian Methods), and Hierarchical Bayesian Models for Spectral Noise Estimation, Supervised vs. Unsupervised Feature Importance Comparison, Classifier Stability Across Bootstrapped Datasets, Mixed-Effects Modeling for Inter-Sample Variability, Spearman/Pearson Correlation Analysis of Replicates, Threshold-Based Peak Area Reproducibility, Dynamic Quality Control Filtering via Anomaly Detection, ROC Curve and AUC-Based Feature Evaluation, False Discovery Rate (FDR) Control in Biomarker Selection, Cross-Validation Performance Metrics (Accuracy, Precision, Recall, F1-score), Autoencoder-Based Feature Reduction & Anomaly Detection.

[000249] Embodiment 25. The method of any of Embodiments 20-24, wherein quality metric is a score, which when normalized from a range of 0-1, is a threshold of at least one of 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 0.96, 0.97, 0.98, and 0.99.

[000250] Embodiment 26. The method of any of Embodiments 20-25, wherein the disease state is one or more pathological state.

[000251] Embodiment 27. The method of any of Embodiments 20-26, wherein the disease state is one or more of kidney cancer, bladder cancer, and prostate cancer.

[000252] Embodiment 28. The method of any of Embodiments 20-27, further comprising:

displaying, a user interface, at least one widget of a dashboard, the widgets display at least one of a diagnosis, a quality control metric, biomarkers ranked above a threshold, a biomarker profile, and a model reasoning widget with a SHAP Waterfall Plot.

[000253] Embodiment 29. The method of any of Embodiments 20-28, wherein the first tensor is a first vector and the second tensor is a second vector.

[000254] Embodiment 30. A method comprising:

receiving, from each of a plurality of machine-learning models, a prediction of a state of a disease state according to the method of Embodiment 18, thereby resulting in a plurality of predictions, the plurality of predictions including at least a first disease state and a second disease state; and

ranking, based on the plurality of predictions, the first disease state and the second disease state.

[000255] Embodiment 31. The method of Embodiment 30, wherein ranking the first disease state and the second disease state includes summing the predictions of each disease state, and ranking the disease states by their summed total.

[000256] Embodiment 32. The method of Embodiment 30, wherein ranking the first disease state and the second disease state includes:

weighting the predictions of each disease state;

summing the weighted predictions of each disease state; and

ranking the disease states by their summed total.

[000257] Embodiment 33. A method comprising:

loading a plurality of disease states, each disease state associated with one of a plurality of subjects;

loading a plurality of tabular data of mass spectrometry reads, each tabular data representing a sample of a plurality of volatile organic compounds (VOCs) extracted from a biological sample from one of the plurality of subjects, the tabular data comprising a plurality of tensors, the tensors including first tensor comprising a mass-to-charge ratio dimension, a second tensor comprising an abundance dimension, and a third tensor comprising a retention time dimension;

for each of the plurality of tabular data:

extracting a first plurality of features from the abundance matrix, each feature of the first plurality associated with one of the VOCs;

generating, for each of the first plurality of features, a quality metric based on the abundance matrix;

selecting a second plurality of features from a set of the first plurality of features that have a quality score above a threshold;

generating a training dataset comprising, for each subject, (a) the disease state associated with that subject and (b) the second plurality of features; and

training a machine-learning model on the training dataset, the resulting machine-learning model configured to receive a third plurality of features and output a prediction of a disease state.

[000258] Embodiment 34. The method of Embodiment 33, wherein extracting the first plurality of features further includes comparing the tabular data to a database of biomarkers, the comparison resulting in the first plurality of features.

[000259] Embodiment 35. The method of any of Embodiments 33-34, wherein the plurality of VOCs include at least one of VOCs having unidentified biomarkers.

[000260] Embodiment 36. The method of any of Embodiments 33-35, wherein the biological sample is a urine sample.

[000261] Embodiment 37. The method of any of Embodiments 33-36, wherein the quality metric is calculated based on one or more of the following methods: Mahalanobis Distance, Gaussian Mixture Model (GMM) Log Likelihood, Kolmogorov-Smirnov (KS) Test, Anderson-Darling Test, T-Test for Mean Differences, Principal Component Analysis (PCA) for Dimensionality Reduction & Visualization, Isolation Forest for Anomaly Detection, Calibration Curve & Logistic Regression, Dynamic Retention Time Termination in GC-MS, Peak Signal-to-Noise Ratio Assessment, Internal Standard Normalization, Feature Selection by Variance Thresholding, Outlier Detection by Robust Z-Score Analysis, Levene's Test for Homogeneity of Variance, Quality Metrics from Spectral Deconvolution (e.g., NIST Library Matching), Mass Spectrometry Data Clustering (K-Means, Hierarchical), Principal Component Regression (PCR), Random Forest Feature Selection, Gradient Boosting Model Evaluation, Variance Explained by Principal Components, SHAP Value Feature Importance for Biomarkers, Bayesian Inference for Outlier Detection, Machine Learning-Based Feature Selection, Peak Alignment and Retention Time Drift Correction, Metabolite Identification Confidence Scoring, Multi-Dimensional Scaling (MDS) for Batch Effects, Weighted Recursive Feature Elimination (WRFE), Euclidean Distance-Based Sample Similarity Scoring, Data Imputation Quality Metrics (kNN, Bayesian Methods), and Hierarchical Bayesian Models for Spectral Noise Estimation, Supervised vs. Unsupervised Feature Importance Comparison,

Classifier Stability Across Bootstrapped Datasets, Mixed-Effects Modeling for Inter-Sample Variability, Spearman/Pearson Correlation Analysis of Replicates, Threshold-Based Peak Area Reproducibility, Dynamic Quality Control Filtering via Anomaly Detection, ROC Curve and AUC-Based Feature Evaluation, False Discovery Rate (FDR) Control in Biomarker Selection, Cross-Validation Performance Metrics (Accuracy, Precision, Recall, F1-score), Autoencoder-Based Feature Reduction & Anomaly Detection.

[000262] Embodiment 38. The method of any of Embodiments 33-37, wherein quality metric is a score, which when normalized from a range of 0-1, is a threshold of at least one of 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 0.96, 0.97, 0.98, and 0.99.

[000263] Embodiment 39. The method of any of Embodiments 33-38, wherein the disease state is one or more of kidney cancer, bladder cancer, and prostate cancer.

[000264] Embodiment 40. The method of any of Embodiments 33-39, the method further comprising:

[000265] displaying, a user interface, at least one widget of a dashboard, the widgets display at least one of a diagnosis, a quality control metric, biomarkers ranked above a threshold, a biomarker profile, and a model reasoning widget with a SHAP Waterfall Plot.

[000266] Embodiment 41. A method of diagnosing a biological condition in a subject by analysis of urine-derived volatile organic compounds (VOCs), the method comprising subjecting VOCs of a urine-derived sample to gas chromatography mass spectrometry (GC-MS), collecting results of GC-MS, and comparing the results to a reference.

[000267] Embodiment 42. The method of embodiment 41, wherein the VOCs are adsorbed on a substrate and desorbed prior to gas chromatography.

[000268] Embodiment 43. The method of embodiment 41 or 42, wherein the GC-MS includes measurement of an internal standard.

[000269] Embodiment 44. The method of embodiment 43, wherein an internal standard is adsorbed on the substrate and desorbed with the VOCs prior to gas chromatography, whereby the internal standard is measured by GC-MS.

[000270] Embodiment 45. The method of any of embodiments 43-44, wherein the internal standard comprises mirex.

[000271] Embodiment 46. The method of any one of embodiments 41-45, wherein the mass spectrometry range of detection is about 20 to about 500 m/z, about 20 to about 550 m/z, or about 20 to about 600 m/z.

[000272] Embodiment 47. The method of any one of embodiments 41-46, wherein the results of GC-MS include the determination of mass-to-charge ratios, signal intensity, and retention time for a plurality of VOCs.

[000273] Embodiment 48. The method of embodiment 47, wherein the plurality of VOCs comprises at least about 100 VOCs, at least about 500 VOCs, or at least about 1000 VOCs.

[000274] Embodiment 49. The method of any one of embodiments 41-48, wherein the comparison to the reference comprises determination of the presence, absence, or level of one or more peaks in the collected results of GC-MS.

[000275] Embodiment 50. The method of any one of embodiments 41-49, wherein the comparison to the reference comprises determining the relative level of peaks present in the collected results of GC-MS by dividing the area of the peaks by the area of the peak of an internal standard, optionally wherein the internal standard is mirex.

[000276] Embodiment 51. The method of any one of embodiments 41-50, wherein one or more of the peaks optionally correspond to a known or predicted VOC.

[000277] Embodiment 52. The method of any one of embodiments 41-51, wherein comparing the results of GC-MS to a reference comprises applying a predictive model to predict a categorical state.

[000278] Embodiment 53. The method of embodiment 52, wherein the predictive model further provides a confidence score for the prediction.

[000279] Embodiment 54. The method of embodiment 53, wherein the method comprises ceasing data collection if, during the collecting of results, the confidence score meets or exceeds a predetermined threshold.

[000280] Embodiment 55. The method of embodiment 54, wherein the predetermined threshold is a percentage value, a numerical value, a categorical value, an ordinal value, a probability estimate (logistic regression, support vector machines, and neural networks), a softmax score (deep learning), a confidence interval, a margin (random forests), a prediction interval, an entropy-based measure, a Bayesian confidence measure (neural networks), a gini coefficient, an impurity measure (random forest), an anomaly score, or a ranking score.

[000281] Embodiment 56. The method of any one of embodiments 41-55 wherein the biological condition is cancer.

[000282] Embodiment 57. The method of any one of embodiments 41-56, further comprising treating a subject determined to have the biological condition.

[000283] Embodiment 58. The method of embodiment 57, wherein biological condition is cancer and the treatment comprises administering at least a chemotherapeutic agent to the subject.

[000284] Embodiment 59. A method for building a classifier that classifies a sample into a categorical state based on volatile organic compounds ("VOCs"), comprising:

(a) providing a plurality of samples classified different categorical states;

(b) analyzing each of the samples by fractionating volatile molecules in the sample by chromatography, and analyzing a plurality of fractions by mass spectrometry, to produce a data set comprising, for each of a plurality of samples, abundance data comprising mass spectra for the plurality of fractions; and

(c) training a machine learning algorithm on the dataset to build a predictive model that classifies a sample into a categorical state.

[000285] Embodiment 60. The method of embodiment 59, wherein the samples are biological samples.

[000286] Embodiment 61. The method of embodiment 60, wherein the biological samples are selected from urine, blood, saliva, feces, sputum, sebum, semen, sweat, breast milk, flatus and exhaled breath.

[000287] Embodiment 62. The method of any one of embodiments 59-61, wherein the categorical states are pathological states.

[000288] Embodiment 63. The method of embodiment 62, wherein pathological states are selected from cancer, infection, autoimmune disease, cardiovascular disease, metabolic disorder, neurological disorder, respiratory disease, gastrointestinal disease, musculoskeletal disorder, psychiatric/behavioral conditions, and dermatological conditions.

[000289] Embodiment 64. The method of any one of embodiments 59-63, wherein the chromatography is selected from gas chromatography, liquid chromatography, supercritical fluid chromatography, parallel comprehensive 2-dimensional gas chromatography, and series comprehensive 2-dimensional gas chromatography.

[000290] Embodiment 65. The method of any one of embodiments 59-64, wherein the mass spectrometry is selected from time-of-flight mass spectrometry, quadrupole mass spectrometry, ion trap mass spectrometry, trap mass spectrometry and quadrupole time-of-flight mass spectrometry, magnetic deflection spectrometry, electrostatic spectrometry, or tandem mass spectrometry.

[000291] Embodiment 66. The method of any one of embodiments 59-65, wherein the data acquisition time for a sample is between one minute and one hour.

[000292] Embodiment 67. The method of any one of embodiments 59-66, wherein the mass spectra are produced at a data acquisition rate of between 1-100 spectra per second.

[000293] Embodiment 68. The method of any one of embodiments 59-67, wherein the abundance data comprise between 2000 and 300,000 mass spectra.

[000294] Embodiment 69. The method of any one of embodiments 59-68, wherein the abundance data further comprise chromatographic data.

[000295] Embodiment 70. The method of any one of embodiments 59-69, wherein the dataset comprises, for each sample, a categorical state.

[000296] Embodiment 71. The method of any one of embodiments 59-70, wherein the machine learning algorithm is selected from a neural network, a support vector machine, principal components analysis, random forest analysis, K-means clustering, Gaussian mixture models, gradient boosting machines, decision trees, logistic regression, naïve bayes, k-nearest neighbors, decision tree, and support vector machine.

[000297] Embodiment 72. The method of any one of embodiments 59-71, wherein the classifier provides a confidence score of the classification.

[000298] Embodiment 73. The method of any one of embodiments 59-72, wherein training the machine learning algorithm is performed on raw feature.

[000299] Embodiment 74. The method of any one of embodiments 59-73, wherein training the machine learning algorithm does not comprise identifying molecular species within the dataset.

[000300] Embodiment 75. A method of classifying a sample into a categorical state comprising:

(a) providing sample;

(b) analyzing the sample by fractionating volatile molecules in the sample by chromatography, and analyzing a plurality of fractions by mass spectrometry, to produce a data set comprising abundance data comprising mass spectra for the plurality of fractions; and

(c) applying a predictive model to the dataset to predict a categorical state.

[000301] Embodiment 76. The method of embodiment 75, wherein the predictive model further provides a confidence score for the prediction.

[000302] Embodiment 77. A method comprising:

(a) performing a first prediction step comprising:

(i) collecting data from a separation step in an analysis of analytes in a sample through a first residence time in the separation system, to produce a dataset;

(ii) applying a predictive model to the dataset to predict a categorical state with a confidence score for the sample;

(iii) determining whether the confidence score meets or exceeds a threshold;

(b) performing a step comprising:

(i) if the confidence score meets or exceeds the threshold, ceasing data collection and reporting the categorical state and confidence score; or

(ii) if the confidence score does not meet or exceed the threshold:

(1) continuing to collect data through a second residence time and adding the data to the dataset;

(2) applying the predictive model to the dataset to predict the categorical state with a confidence score for the sample;

(3) determining whether the confidence score meets or exceeds the threshold; and

(c) repeating step (d) until the confidence score meets or exceeds the threshold or there is a plateau in the improvement of the confidence score.

[000303] Embodiment 78. A method comprising:

(a) performing a first prediction step comprising:

(i) collecting abundance data from a gas chromatography-mass spectrometry ("GC-MS") run on a sample, wherein the abundance data comprises a plurality of mass spectra through a first retention time, to produce a dataset;

(ii) applying a predictive model to the dataset to predict a categorical state with a confidence score for the sample;

(iii) determining whether the confidence score meets or exceeds a threshold;

(b) performing a step comprising:

(i) if the confidence score meets or exceeds the threshold, ceasing data collection and reporting the categorical state and confidence score; or

(ii) if the confidence score does not meet or exceed the threshold:

(1) continuing to collect abundance data through a second retention time and adding the abundance data to the dataset;

(2) applying the predictive model to the dataset to predict the categorical state with a confidence score for the sample;

(3) determining whether the confidence score meets or exceeds the threshold; and

(c) repeating step (b) until the confidence score meets or exceeds the threshold or there is a plateau in the improvement of the confidence score.

[000304] Embodiment 79. The method of embodiment 78, wherein the confidence score threshold is a percentage value, a numerical value, a categorical value, an ordinal value, a probability estimate (logistic regression, support vector machines, and neural networks), a softmax score (deep learning), a confidence interval, a margin (random forests), a prediction interval, an entropy-based measure, a Bayesian confidence measure (neural networks), a gini coefficient, an impurity measure (random forest), an anomaly score, or a ranking score.

[000305] Embodiment 80. The method of embodiment 78-79, wherein the first retention time is no more than 1 minute, 5 minutes, 10 minutes, 15 minutes, 20 minutes, 30 minutes, or 40 minutes.

[000306] Embodiment 81. The method of any one of embodiments 78-80, wherein the second retention times are increments of no more than 1 minute, 5 minutes, 10 minutes or 15 minutes.

[000307] Embodiment 82. The method of any one of embodiments 78-81, wherein the second retention time immediately after the first retention time is no more than 5 minutes.

[000308] Embodiment 83. The method of any one of embodiments 78-82, wherein the improvement in the confidence score plateaus, if the confidence score increases by less than 5% over 10% of the elapsed retention time.

[000309] Embodiment 84. The method of any one of embodiments 78-83, wherein when the confidence score is a member of an ordinal set comprising low, medium, and high confidence, step (b) is repeated until the confidence score satisfies the medium confidence level.

[000310] Embodiment 85. The method of any one of embodiments 78-84, wherein the improvement plateaus when the rate of improvement of the confidence score of collected data drops below a rate threshold.

[000311] Embodiment 86. The method of any one of embodiments 78-85, wherein the improvement plateaus when the predicted rate of improvement of the confidence score for future collected data drops below a rate.

[000312] Embodiment 87. A method of treating a subject comprising:

(a) diagnosing a subject as having a pathological condition by the method of any of embodiments 78-86; and

(b) administering the therapeutic intervention effective to treat the condition to the subject.

[000313] Embodiment 88. The method of any one of embodiments 41-58, wherein the method comprises the method of any one of embodiments 1-40.

[000314] Embodiment 89. The method of any one of embodiments 41-58, wherein the method comprises the method of any one of embodiments 59-87.

[000315] Embodiment 90. A system comprising:

a computing node comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor of the computing node to cause the processor to perform the method of any of Embodiments 1-89.

[000316] Embodiment 91. A computer program product for predicting a disease state, the computer program product comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor to cause the processor to perform the method of any of Embodiments 1-89.

EXAMPLES

[000317] The present disclosure and Examples demonstrate methods and compositions provided herein, as well as their advantageous use for the detected, diagnosed, and/or monitored of biological conditions such as cancer.

Example 1: Analysis of volatile organic compounds (VOCs) in urine.

[000318] Urine samples of 10-15 mL urine are collected from subjects in sterile containers following standardized protocols to ensure consistency and reliability of the VOC analysis. Prior to VOC extraction, urine samples are centrifuged for 10 minutes at 300 g. From each sample, 1 mL of urine is diluted to a final volume of 20 mL with HPLC-grade water and then mixed with 300 uL of 1 ppm Mirex solution (in methanol; Dr. Ehrenstorfer GmbH, Germany) as a standard for relative concentration, and 600 uL of 2 M hydrochloric acid. Urinary VOCs are extracted using Stir Bar Sorptive Extraction (SBSE) by stirring the diluted, final sample with a Twister™ stir bar for 2 hours at 1000 rpm. The VOCs in the urine sample are consequently adsorbed by the polydimethylsiloxane (PDMS) of the stir bar.

[000319] After VOC extraction, the stir bar is desorbed thermally by holding the stir bar at 45 °C for 0.5 min, subsequently heating the stir bar at 300 °C for 5 min. Urinary VOCs in the desorption gas flow are concentrated in a cold injection system at -40 °C.

[000320] Urinary VOC samples are surveyed with GC-MS. Briefly, column oven temperature is held at 35 °C for 5 min, and ramped to 300 °C at 10 °C/min and subsequently held for 10 min. Mass spectrometry is performed in scan mode with a window of 20-500 m/z. Detected features are identified with the National Institute of Standards and Technology Library (NIST17) and analyzed to assign a categorical state (e.g., state of bladder cancer) to the associated urine sample.

[000321] Thus, in one exemplary approach, urine samples are collected in sterile containers following standardized protocols to ensure consistency and reliability of the VOC analysis. Urinary volatile organic compounds (VOCs) are extracted using Stir Bar Sorptive Extraction (SBSE). This technique involves immersing a stir bar coated with a polydimethylsiloxane (PDMS) layer into urine samples, where it absorbs VOCs. After extraction, the stir bar is then desorbed either thermally or through solvent desorption for subsequent analysis by gas

chromatography-mass spectrometry (GC-MS). SBSE is selected for its efficiency in concentrating VOCs from complex biological matrices like urine, significantly enhancing the sensitivity and specificity of the cancer diagnosis method.

Example 2: Prediction of Bladder Cancer Using Urinary VOCs

[000322] The present Example demonstrates in a representative trial study that urinary VOCs can be used in the diagnosis of bladder cancer (BCa).

[000323] A study was performed with urine samples from 258 human subjects, including 58 with biopsy-proven BCa and 200 healthy individuals, to explore the utility of metabolomics in BCa detection and grading. Urine samples from subject were collected and analyzed as described in Example 1. More than 20,000 features or more than one biomarkers were identified from the data set as correlated with BCa. Exemplary illustrations of features, solely for visualization of relevant principles, are provided in the GC chromatogram of Figure 2, exemplary mass spectrum of Figure 3, and exemplary abundance data matrix of Figures 4 and 5. An area under the curve (AUC) of 0.98 was generated from the present analysis (data not shown), demonstrating the high performance of the test in predicting BCa from urinary samples.

[000324] These data confirm that the classifier can predict BCa from urinary samples. These data further signal the utility of the classification method and the VOC-based detection of BCa in enhancing diagnosis, treatment efficacy, and biopsy targeting, and mitigating the overdiagnosis and overtreatment issues associated with current screening methods based on identification of biomarkers.

[000325] Thus, in one exemplary study, in a significant advancement towards improving bladder cancer (BCa) diagnostics, a study analyzed urine samples from 258 human subjects, including 58 with biopsy-proven BCa and 200 healthy individuals, to explore the utility of metabolomics in BCa detection. Employing stir bar sorptive extraction and gas chromatography-mass spectrometry, more than 20,000 features or more than one biomarkers were identified that were correlated with bladder cancer. Some or all of these features can be used with the classifier. The classifier generated produces an AUC of 0.98. One aim is to enhance diagnosis, treatment efficacy, and biopsy targeting, and mitigate the overdiagnosis and overtreatment issues associated with current screening methods.

Example 3: Machine-Learning Methods to Predict Disease State Using Biological Samples

[000326] The present Example demonstrates a representative use case of the machine-learning models as described herein.

[000327] A biological sample, such as a urine sample, can be collected from a subject. In some examples, the collection can occur at a medical facility or laboratory. The collection can occur in other locations, and a medical facility or laboratory can receive the collected sample and forward it for mass spectrometry analysis or perform the mass spectrometry analysis themselves. The urine sample can be analyzed by a mass spectrometer, the analysis resulting in an abundance matrix.

[000328] The abundance matrix can then be loaded by an analysis facility. The analysis facility can run the machine-learning models described above, either on site or with a SaaS model. The abundance matrix can be provided to the trained machine-learning models, which can be executed by local processors or by cloud computing server(s). The machine-learning models can thereby provide likelihoods of one ore more pathology states.

[000329] In some examples, the abundance matrix can be provided to multiple models in the same manner as described above. For example, each model can be configured to provide a probability of one disease state, pathological state, *etc*. Outputs from one model can be used to provide likelihood of that disease state. The models can be used in combination to provide a likelihoods of multiple disease states, including providing a most likely disease states. In some embodiments, one model can be configured to provide multiple disease likelihoods or pathological states.

[000330] The likelihoods and other metrics provided by the model can then be displayed to an end-user, such as a patient or a medical professional, in a dashboard illustrating the disease likelihood, most likely disease state, a ranking of disease state, or other graphical representation. The disease likelihoods can then be used by the medical professional or the patient to order further tests to verify

Example 4: Training Machine-Learning to Predict Disease State Using Biological Sample

[000331] The present Example demonstrates a representative use case of training the machine-learning models as described herein.

[000332] A training dataset comprising mass spectrometry data associated with disease state of subject's data for that data is provided. The data can be loaded by a database, or developed via sample collection and other pathology/disease state detection methods to determine the ground truth. This training set can then be provided to a model training factory. The model training factory can be a plurality of servers either locally or in the cloud executing a model training method. The model training factory then can output the resulting trained model.

Example 5: Prediction of Prostate Cancer Using Urinary VOCs

[000333] The present Example demonstrates in a representative trial study that urinary VOCs can be used in the diagnosis of prostate cancers (PCa).

[000334] A study was performed with urine samples from 227 men, including 27 with biopsy-proven PCa and 200 healthy individuals, to explore the utility of metabolomics in PCa detection and grading. Urine samples from subject were collected and analyzed as described in Example 1. More than 20,000 features or more than one biomarkers were identified from the data set as correlated with PCa. Exemplary illustrations of features, solely for visualization of relevant principles, are provided in the GC chromatogram of **Figure 2**, exemplary mass spectrum of **Figure 3**, and exemplary abundance data matrix of **Figures 4** and **Figure 5**. An area under the curve (AUC) of 0.97 was generated from the present analysis (**Figure 13**), demonstrating the high performance of the test in predicting PCa from urinary samples.

[000335] These data confirm that the classifier can predict PCa from urinary samples and distinguish between low-grade and intermediate/high-grade PCa. These data further signal the utility of the classification method and the VOC-based detection of PCa in enhancing diagnosis, treatment efficacy, and biopsy targeting, and mitigating the overdiagnosis and overtreatment issues associated with current screening methods based on identification of biomarkers.

[000336] Thus, in one exemplary study, in a significant advancement towards improving prostate cancer (PCa) diagnostics, a study analyzed urine samples from 227 men, including 27

with biopsy-proven PCa and 200 healthy individuals, to explore the utility of metabolomics in PCa detection and grading. Employing stir bar sorptive extraction and gas chromatography-mass spectrometry, more than 20,000 features or more than one biomarkers were identified that were correlated with prostate cancer. Some or all of these features can be used with the classifier. The classifier generated produces an AUC of 0.97 (See Figure 6). The classifier can distinguish between low-grade and intermediate/high-grade PCa. One aim is to enhance diagnosis, treatment efficacy, and biopsy targeting, and mitigate the overdiagnosis and overtreatment issues associated with current screening methods.

[000337] As used herein, the following meanings apply unless otherwise specified. The words "can" and "may" are used in a permissive sense (i.e., meaning having the potential to), rather than the mandatory sense (i.e., meaning must). The words "include", "including", and "includes" and the like mean including, but not limited to. The singular forms "a," "an," and "the" include plural referents. Thus, for example, reference to "an element" includes a combination of two or more elements, notwithstanding use of other terms and phrases for one or more elements, such as "one or more." The phrase "at least one" includes "one", "one or more", "one or a plurality", and, therefore, contemplates the use of the term "a plurality". The term "or" is, unless indicated otherwise, non-exclusive, i.e., encompassing both "and" and "or." The term "any of" between a modifier and a sequence means that the modifier modifies each member of the sequence. So, for example, the phrase "at least any of 1, 2 or 3" means "at least 1, at least 2 or at least 3". The term "about" refers to a range that is 5% plus or minus from a stated numerical value within the context of the particular usage. The term "consisting essentially of" refers to the inclusion of recited elements and other elements that do not materially affect the basic and novel characteristics of a claimed combination.

OTHER EMBODIMENTS

[000338] It should be understood that the description and the drawings are not intended to limit the invention to the particular form disclosed, but to the contrary, the intention is to cover all modifications, equivalents, and alternatives falling within the spirit and scope of the present invention as defined by the appended claims. Further modifications and alternative embodiments of various aspects of the invention will be apparent to those skilled in the art in view of this description. Accordingly, this description and the drawings are to be construed as illustrative only and are for the purpose of teaching those skilled in the art the general manner of carrying out the invention. It is to be understood that the forms of the invention shown and described herein are to be taken as examples of embodiments. Elements and materials may be substituted for those illustrated and described herein, parts and processes may be reversed or omitted, and certain features of the invention may be utilized independently, all as would be apparent to one skilled in the art after having the benefit of this description of the invention. Changes may be made in the elements described herein without departing from the spirit and scope of the invention as described in the following claims.

[000339] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

CLAIMS

What is claimed is:

1. A method comprising:

loading an abundance matrix that represents mass spectrometry reads of a sample of a plurality of volatile organic compounds (VOCs) extracted from a biological sample, the abundance matrix representing each of the plurality of VOCs in a mass-to-charge ratio dimension, an abundance dimension, and a retention time dimension;

providing the abundance matrix to a machine-learning model, the machine-learning model trained on abundance matrixes of biomarkers of VOCs collected from healthy and diseased subjects; and

receiving, from the machine-learning model, a prediction of a state of a disease state.

2. The method of Claim 1, wherein the plurality of VOCs include at least one of VOCs having unidentified biomarkers.

3. The method of any of Claims 1-2, wherein the biological sample is a urine sample.

4. The method of any of Claims 1-3, wherein the disease state is a pathological state.

5. The method of any of Claims 1-4, wherein the disease state is one or more of kidney cancer, bladder cancers, and prostate cancer.

6. The method of any of Claims 1-5, further comprising:

displaying, a user interface, at least one widget of a dashboard, the widgets display at least one of a diagnosis, a quality control metric, biomarkers ranked above a threshold, a biomarker profile, and a model reasoning widget with a SHAP Waterfall Plot.

7. The method of any of Claims 1-6, further comprising:

generating the abundance matrix based on raw data received from a mass spectrometer.

8. A method comprising:

receiving, from each of a plurality of machine-learning models, a prediction of a state of a disease state according to the method of Claim 1, thereby resulting in a plurality of predictions, the plurality of predictions including at least a first disease state and a second disease state; and

ranking, based on the plurality of predictions, the first disease state and the second disease state.

9. The method of Claim 8, wherein ranking the first disease state and the second disease state includes summing the predictions of each disease state, and ranking the disease states by their summed total.

10. The method of Claim 8, wherein ranking the first disease state and the second disease state includes:

weighting the predictions of each disease state;

summing the weighted predictions of each disease state; and

ranking the disease states by their summed total.

11. A method comprising:

loading a plurality of disease states, each disease state associated with one of a plurality of subjects;

loading a plurality of abundance matrices, each abundance matrix representing mass spectrometry reads of a sample of a plurality of volatile organic compounds (VOCs) extracted from a biological sample from one of the plurality of subjects, the mass spectrometry read representing each of the plurality of VOCs, each abundance matrix representing each of the plurality of VOCs in a mass-to-charge ratio dimension, an abundance dimension, and a retention time dimension; and

training a machine-learning model on the plurality of abundance matrices and the plurality of disease states, the resulting machine-learning model configured to receive a second abundance matrix and output a prediction of a disease state.

12. The method of Claim 11, wherein the plurality of VOCs include at least one of VOCs having unidentified biomarkers.

13. The method of any of Claims 11-12, wherein the biological sample is a urine sample.

14. The method of any of Claims 11-13, wherein the disease state is a pathological state.

15. The method of any of Claims 11-14 wherein the disease state is one or more of kidney cancer, bladder cancer, and prostate cancer.

16. The method of any of Claims 11-15, further comprising:

generating the abundance matrix based on raw data received from a mass spectrometer.

17. A method comprising:

generating a first ranked list of biomarkers, the biomarkers ranked by an importance metric;

selecting a plurality of biomarkers from the ranked list, the plurality of biomarkers being the highest ranked biomarkers of the ranked list;

generating a second ranked list of biomarkers, the second ranked list comprising the biomarkers of the first ranked list with the plurality of biomarkers removed; and

training a machine learning model using the method of any of Claims 11-16, wherein the second plurality of biomarkers in the training dataset is the second ranked list of biomarkers.

18. A method comprising:

generating a ranked list of biomarkers, the biomarkers ranked by an importance metric;

selecting a plurality of biomarkers from the ranked list, the plurality of biomarkers being the highest ranked biomarkers of the ranked list; and

training a machine learning model using the method of any of Claims 11-16, wherein the second plurality of biomarkers in the training dataset is the plurality of biomarkers.

19. A method comprising:

generating a first ranked list of biomarkers, the biomarkers ranked by an importance metric;

selecting a plurality of biomarkers from the ranked list, the plurality of biomarkers being the highest ranked biomarkers of the ranked list;

generating a second ranked list of biomarkers, the second ranked list comprising the biomarkers of the first ranked list with the plurality of biomarkers removed;

generating a third ranked list of biomarkers, the third ranked list comprising a plurality of top ranked biomarkers from the second ranked list; and

training a machine learning model using the method of any of Claims 11-16, wherein the second plurality of biomarkers in the training dataset is the third ranked list of biomarkers.

20. A method comprising:

loading tabular data of mass spectrometry reads that represent a sample of a plurality of volatile organic compounds (VOCs) extracted from a biological sample, the mass spectrometry read representing each of the plurality of VOCs, the tabular data comprising a plurality of tensors, the tensors including first tensor comprising a mass-to-charge ratio dimension, a second tensor comprising an abundance dimension, and a third tensor comprising a retention time dimension;

extracting a first plurality of features from the tabular data, each feature of the first plurality associated with one of the VOCs;

generating, for each of the first plurality of features, a quality metric based on the abundance matrix;

selecting a second plurality of features from a set of the first plurality of features that have a quality score above a threshold;

providing the second plurality of features to a machine-learning model, the machine-learning model trained on features of VOCs collected from healthy and diseased subjects; and

receiving, from the machine-learning model, a prediction of a state of a disease state.

21. The method of Claim 20, wherein extracting the first plurality of features further includes comparing the tabular data to a database of features, the comparison resulting in the first plurality of features.

22. The method of any of Claims 20-21, wherein the plurality of VOCs include at least one of VOCs having unidentified biomarkers.

23. The method of any of Claims 20-22, wherein the biological sample is a urine sample.

24. The method of any of Claims 20-23, wherein the quality metric is calculated based on one or more of the following methods: Mahalanobis Distance, Gaussian Mixture Model (GMM) Log Likelihood, Kolmogorov-Smirnov (KS) Test, Anderson-Darling Test, T-Test for Mean Differences, Principal Component Analysis (PCA) for Dimensionality Reduction & Visualization, Isolation Forest for Anomaly Detection, Calibration Curve & Logistic Regression, Dynamic Retention Time Termination in GC-MS, Peak Signal-to-Noise Ratio Assessment, Internal Standard Normalization, Feature Selection by Variance Thresholding, Outlier Detection by Robust Z-Score Analysis, Levene's Test for Homogeneity of Variance, Quality Metrics from Spectral Deconvolution (e.g., NIST Library Matching), Mass Spectrometry Data Clustering (K-Means, Hierarchical), Principal Component Regression (PCR), Random Forest Feature Selection, Gradient

Boosting Model Evaluation, Variance Explained by Principal Components, SHAP Value Feature Importance for Biomarkers, Bayesian Inference for Outlier Detection, Machine Learning-Based Feature Selection, Peak Alignment and Retention Time Drift Correction, Metabolite Identification Confidence Scoring, Multi-Dimensional Scaling (MDS) for Batch Effects, Weighted Recursive Feature Elimination (WRFE), Euclidean Distance-Based Sample Similarity Scoring, Data Imputation Quality Metrics (kNN, Bayesian Methods), and Hierarchical Bayesian Models for Spectral Noise Estimation, Supervised vs. Unsupervised Feature Importance Comparison, Classifier Stability Across Bootstrapped Datasets, Mixed-Effects Modeling for Inter-Sample Variability, Spearman/Pearson Correlation Analysis of Replicates, Threshold-Based Peak Area Reproducibility, Dynamic Quality Control Filtering via Anomaly Detection, ROC Curve and AUC-Based Feature Evaluation, False Discovery Rate (FDR) Control in Biomarker Selection, Cross-Validation Performance Metrics (Accuracy, Precision, Recall, F1-score), Autoencoder-Based Feature Reduction & Anomaly Detection.

25. The method of any of Claims 20-24, wherein quality metric is a score, which when normalized from a range of 0-1, is a threshold of at least one of 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 0.96, 0.97, 0.98, and 0.99.

26. The method of any of Claims 20-25, wherein the disease state is one or more pathological state.

27. The method of any of Claims 20-26, wherein the disease state is one or more of kidney cancer, bladder cancer, and prostate cancer.

28. The method of any of Claims 20-27, further comprising:

displaying, a user interface, at least one widget of a dashboard, the widgets display at least one of a diagnosis, a quality control metric, biomarkers ranked above a threshold, a biomarker profile, and a model reasoning widget with a SHAP Waterfall Plot.

29. The method of any of Claims 20-28, wherein the first tensor is a first vector and the second tensor is a second vector.

30. A method comprising:

receiving, from each of a plurality of machine-learning models, a prediction of a state of a disease state according to the method of Claim 18, thereby resulting in a plurality of predictions, the plurality of predictions including at least a first disease state and a second disease state; and

ranking, based on the plurality of predictions, the first disease state and the second disease state.

31. The method of Claim 30, wherein ranking the first disease state and the second disease state includes summing the predictions of each disease state, and ranking the disease states by their summed total.

32. The method of Claim 30, wherein ranking the first disease state and the second disease state includes:
weighting the predictions of each disease state;
summing the weighted predictions of each disease state; and
ranking the disease states by their summed total.

33. A method comprising:

loading a plurality of disease states, each disease state associated with one of a plurality of subjects;

loading a plurality of tabular data of mass spectrometry reads, each tabular data representing a sample of a plurality of volatile organic compounds (VOCs) extracted from a biological sample from one of the plurality of subjects, the tabular data comprising a plurality of tensors, the tensors including first tensor comprising a mass-to-charge ratio dimension, a second tensor comprising an abundance dimension, and a third tensor comprising a retention time dimension;

for each of the plurality of tabular data:

extracting a first plurality of features from the abundance matrix, each feature of the first plurality of features associated with one of the VOCs;

generating, for each of the first plurality of features, a quality metric based on the abundance matrix;

selecting a second plurality of features from a set of the first plurality of features that have a quality score above a threshold;

generating a training dataset comprising, for each subject, (a) the disease state associated with that subject and (b) the second plurality of features; and

training a machine-learning model on the training dataset, the resulting machine-learning model configured to receive a third plurality of features and output a prediction of a disease state.

34. The method of Claim 33, wherein extracting the first plurality of features further includes comparing the tabular data to a database of biomarkers, the comparison resulting in the first plurality of features.

35. The method of any of Claims 33-34, wherein the plurality of VOCs include at least one of VOCs having unidentified biomarkers.

36. The method of any of Claims 33-35, wherein the biological sample is a urine sample.

37. The method of any of Claims 33-36, wherein the quality metric is calculated based on one or more of the following methods: Mahalanobis Distance, Gaussian Mixture Model (GMM) Log Likelihood, Kolmogorov-Smirnov (KS) Test, Anderson-Darling Test, T-Test for Mean Differences, Principal Component Analysis (PCA) for Dimensionality Reduction & Visualization, Isolation Forest for Anomaly Detection, Calibration Curve & Logistic Regression, Dynamic Retention Time Termination in GC-MS, Peak Signal-to-Noise Ratio Assessment, Internal Standard Normalization, Feature Selection by Variance Thresholding, Outlier Detection by Robust Z-Score Analysis, Levene's Test for

Homogeneity of Variance, Quality Metrics from Spectral Deconvolution (e.g., NIST Library Matching), Mass Spectrometry Data Clustering (K-Means, Hierarchical), Principal Component Regression (PCR), Random Forest Feature Selection, Gradient Boosting Model Evaluation, Variance Explained by Principal Components, SHAP Value Feature Importance for Biomarkers, Bayesian Inference for Outlier Detection, Machine Learning-Based Feature Selection, Peak Alignment and Retention Time Drift Correction, Metabolite Identification Confidence Scoring, Multi-Dimensional Scaling (MDS) for Batch Effects, Weighted Recursive Feature Elimination (WRFE), Euclidean Distance-Based Sample Similarity Scoring, Data Imputation Quality Metrics (kNN, Bayesian Methods), and Hierarchical Bayesian Models for Spectral Noise Estimation, Supervised vs. Unsupervised Feature Importance Comparison, Classifier Stability Across Bootstrapped Datasets, Mixed-Effects Modeling for Inter-Sample Variability, Spearman/Pearson Correlation Analysis of Replicates, Threshold-Based Peak Area Reproducibility, Dynamic Quality Control Filtering via Anomaly Detection, ROC Curve and AUC-Based Feature Evaluation, False Discovery Rate (FDR) Control in Biomarker Selection, Cross-Validation Performance Metrics (Accuracy, Precision, Recall, F1-score), Autoencoder-Based Feature Reduction & Anomaly Detection.

38. The method of any of Claims 33-37, wherein quality metric is a score, which when normalized from a range of 0-1, is a threshold of at least one of 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 0.96, 0.97, 0.98, and 0.99.

39. The method of any of Claims 33-38, wherein the disease state is one or more of kidney cancer, bladder cancer, and prostate cancer.

40. The method of any of Claims 33-39, the method further comprising:

displaying, a user interface, at least one widget of a dashboard, the widgets display at least one of a diagnosis, a quality control metric, biomarkers ranked above a threshold, a biomarker profile, and a model reasoning widget with a SHAP Waterfall Plot.

41. A method of diagnosing a biological condition in a subject by analysis of urine-derived volatile organic compounds (VOCs), the method comprising subjecting VOCs of a urine-derived sample to gas chromatography mass spectrometry (GC-MS), collecting results of GC-MS, and comparing the results to a reference.

42. The method of claim 41, wherein the VOCs are adsorbed on a substrate and desorbed prior to gas chromatography.

43. The method of claim 41 or 42, wherein the GC-MS includes measurement of an internal standard.

44. The method of claim 43, wherein an internal standard is adsorbed on the substrate and desorbed with the VOCs prior to gas chromatography, whereby the internal standard is measured by GC-MS.

45. The method of any of claims 43-44, wherein the internal standard comprises mirex.

46. The method of any one of claims 41-45, wherein the mass spectrometry range of detection is about 20 to about 500 m/z, about 20 to about 550 m/z, or about 20 to about 600 m/z.

47. The method of any one of claims 41-46, wherein the results of GC-MS include the determination of mass-to-charge ratios, signal intensity, and retention time for a plurality of VOCs.

48. The method of claim 47, wherein the plurality of VOCs comprises at least about 100 VOCs, at least about 500 VOCs, or at least about 1000 VOCs.

49. The method of any one of claims 41-48, wherein the comparison to the reference comprises determination of the presence, absence, or level of one or more peaks in the collected results of GC-MS.

50. The method of any one of claims 41-49, wherein the comparison to the reference comprises determining the relative level of peaks present in the collected results of GC-MS by dividing the area of the peaks by the area of the peak of an internal standard, optionally wherein the internal standard is mirex.

51. The method of any one of claims 41-50, wherein one or more of the peaks optionally correspond to a known or predicted VOC.

52. The method of any one of claims 41-51, wherein comparing the results of GC-MS to a reference comprises applying a predictive model to predict a categorical state.

53. The method of claim 52, wherein the predictive model further provides a confidence score for the prediction.

54. The method of claim 53, wherein the method comprises ceasing data collection if, during the collecting of results, the confidence score meets or exceeds a predetermined threshold.

55. The method of claim 54, wherein the predetermined threshold is a percentage value, a numerical value, a categorical value, an ordinal value, a probability estimate (logistic regression, support vector machines, and neural networks), a softmax score (deep learning), a confidence interval, a margin (random forests), a prediction interval, an entropy-based measure, a Bayesian confidence measure (neural networks), a gini coefficient, an impurity measure (random forest), an anomaly score, or a ranking score.

56. The method of any one of claims 41-55 wherein the biological condition is cancer.

57. The method of any one of claims 41-56, further comprising treating a subject determined to have the biological condition.

58. The method of claim 57, wherein biological condition is cancer and the treatment comprises administering at least a chemotherapeutic agent to the subject.

59. A method for building a classifier that classifies a sample into a categorical state based on volatile organic compounds ("VOCs"), comprising:

(a) providing a plurality of samples classified different categorical states;

(b) analyzing each of the samples by fractionating volatile molecules in the sample by chromatography, and analyzing a plurality of fractions by mass spectrometry, to produce a data set comprising, for each of a plurality of samples, abundance data comprising mass spectra for the plurality of fractions; and

(c) training a machine learning algorithm on the dataset to build a predictive model that classifies a sample into a categorical state.

60. The method of claim 59, wherein the samples are biological samples.

61. The method of claim 60, wherein the biological samples are selected from urine, blood, saliva, feces, sputum, sebum, semen, sweat, breast milk, flatus and exhaled breath.

62. The method of any one of claims 59-61, wherein the categorical states are pathological states.

63. The method of claim 62, wherein pathological states are selected from cancer, infection, autoimmune disease, cardiovascular disease, metabolic disorder, neurological disorder, respiratory disease, gastrointestinal disease, musculoskeletal disorder, psychiatric/behavioral conditions, and dermatological conditions.

64. The method of any one of claims 59-63, wherein the chromatography is selected from gas chromatography, liquid chromatography, supercritical fluid chromatography, parallel comprehensive 2-dimensional gas chromatography, and series comprehensive 2-dimensional gas chromatography.

65. The method of any one of claims 59-64, wherein the mass spectrometry is selected from time-of-flight mass spectrometry, quadrupole mass spectrometry, ion trap mass spectrometry, trap mass spectrometry and quadrupole time-of-flight mass spectrometry, magnetic deflection spectrometry, electrostatic spectrometry, or tandem mass spectrometry.

66. The method of any one of claims 59-65, wherein the data acquisition time for a sample is between one minute and one hour.

67. The method of any one of claims 59-66, wherein the mass spectra are produced at a data acquisition rate of between 1-100 spectra per second.

68. The method of any one of claims 59-67, wherein the abundance data comprise between 2000 and 300,000 mass spectra.

69. The method of any one of claims 59-68, wherein the abundance data further comprise chromatographic data.

70. The method of any one of claims 59-69, wherein the dataset comprises, for each sample, a categorical state.

71. The method of any one of claims 59-70, wherein the machine learning algorithm is selected from a neural network, a support vector machine, principal components analysis, random forest analysis, K-means clustering, Gaussian mixture models, gradient boosting

machines, decision trees, logistic regression, naïve bayes, k-nearest neighbors, decision tree, and support vector machine.

72. The method of any one of claims 59-71, wherein the classifier provides a confidence score of the classification.

73. The method of any one of claims 59-72, wherein training the machine learning algorithm is performed on raw feature.

74. The method of any one of claims 59-73, wherein training the machine learning algorithm does not comprise identifying molecular species within the dataset.

75. A method of classifying a sample into a categorical state comprising:
(a) providing sample;
(b) analyzing the sample by fractionating volatile molecules in the sample by chromatography, and analyzing a plurality of fractions by mass spectrometry, to produce a data set comprising abundance data comprising mass spectra for the plurality of fractions; and
(c) applying a predictive model to the dataset to predict a categorical state.

76. The method of claim 75, wherein the predictive model further provides a confidence score for the prediction.

77. A method comprising:
(a) performing a first prediction step comprising:
(i) collecting data from a separation step in an analysis of analytes in a sample through a first residence time in the separation system, to produce a dataset;
(ii) applying a predictive model to the dataset to predict a categorical state with a confidence score for the sample;

(iii) determining whether the confidence score meets or exceeds a threshold;

(b) performing a step comprising:

(i) if the confidence score meets or exceeds the threshold, ceasing data collection and reporting the categorical state and confidence score; or

(ii) if the confidence score does not meet or exceed the threshold:

(1) continuing to collect data through a second residence time and adding the data to the dataset;

(2) applying the predictive model to the dataset to predict the categorical state with a confidence score for the sample;

(3) determining whether the confidence score meets or exceeds the threshold; and

(c) repeating step (d) until the confidence score meets or exceeds the threshold or there is a plateau in the improvement of the confidence score.

78. A method comprising:

(a) performing a first prediction step comprising:

(i) collecting abundance data from a gas chromatography-mass spectrometry ("GC-MS") run on a sample, wherein the abundance data comprises a plurality of mass spectra through a first retention time, to produce a dataset;

(ii) applying a predictive model to the dataset to predict a categorical state with a confidence score for the sample;

(iii) determining whether the confidence score meets or exceeds a threshold;

(b) performing a step comprising:

(i) if the confidence score meets or exceeds the threshold, ceasing data collection and reporting the categorical state and confidence score; or

(ii) if the confidence score does not meet or exceed the threshold:

(1) continuing to collect abundance data through a second retention time and adding the abundance data to the dataset;

(2) applying the predictive model to the dataset to predict the categorical state with a confidence score for the sample;

(3) determining whether the confidence score meets or exceeds the threshold; and

(c) repeating step (b) until the confidence score meets or exceeds the threshold or there is a plateau in the improvement of the confidence score.

79. The method of claim 78, wherein the confidence score threshold is a percentage value, a numerical value, a categorical value, an ordinal value, a probability estimate (logistic regression, support vector machines, and neural networks), a softmax score (deep learning), a confidence interval, a margin (random forests), a prediction interval, an entropy-based measure, a Bayesian confidence measure (neural networks), a gini coefficient, an impurity measure (random forest), an anomaly score, or a ranking score.

80. The method of claim 78-79, wherein the first retention time is no more than 1 minute, 5 minutes, 10 minutes, 15 minutes, 20 minutes, 30 minutes, or 40 minutes.

81. The method of any one of claims 78-80, wherein the second retention times are increments of no more than 1 minute, 5 minutes, 10 minutes or 15 minutes.

82. The method of any one of claims 78-81, wherein the second retention time immediately after the first retention time is no more than 5 minutes.

83. The method of any one of claims 78-82, wherein the improvement in the confidence score plateaus, if the confidence score increases by less than 5% over 10% of the elapsed retention time.

84. The method of any one of claims 78-83, wherein when the confidence score is a member of an ordinal set comprising low, medium, and high confidence, step (b) is repeated until the confidence score satisfies the medium confidence level.

85. The method of any one of claims 78-84, wherein the improvement plateaus when the rate of improvement of the confidence score of collected data drops below a rate threshold.

86. The method of any one of claims 78-85, wherein the improvement plateaus when the predicted rate of improvement of the confidence score for future collected data drops below a rate.

87. A method of treating a subject comprising:

(a) diagnosing a subject as having a pathological condition by the method of any of claims 78-86; and

(b) administering the therapeutic intervention effective to treat the condition to the subject.

88. The method of any one of claims 41-58, wherein the method comprises the method of any one of claims 1-40.

89. The method of any one of claims 41-58, wherein the method comprises the method of any one of claims 59-87.

90. A system comprising:

a computing node comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor of the computing node to cause the processor to perform the method of any of Claims 1-89.

91. A computer program product for predicting a disease state, the computer program product comprising a computer readable storage medium having program instructions

embodied therewith, the program instructions executable by a processor to cause the processor to perform the method of any of Claims 1-89.