"Assisted analysis of biological sample data"

Cross-Reference to Related Applications

[0001] The present application claims priority from Australian Provisional Patent Application No 2024900255 filed on 5 February 2024, the contents of which are incorporated herein by reference in their entirety.

Technical Field

[0002] This disclosure relates to user-assisted analysis of biological sample data.

Background

[0003] Measurement technologies for biological samples have been advancing at a rapid pace. For example, nucleic acid sequencing, which is also a measurement of a biological sample, has improved over the last ten years to a point where an entire genome, an entire transcriptome or other sequencing data can be obtained in a clinical setting. Such data provides sequencing information for many genes or even all genes, which is in contrast to gene panel tests, which only test for a small number of genes. However, while these new sequencing methods are available, current computer technology is not able to interpret the sequencing data meaningfully and provide a descriptive analysis. Especially in cases where multiple genes interact – in which case it is an advantage to have the data across all genes – it is difficult for current computer technology to generate meaningful analyses. For other measurements, such as titres of a sample group of individuals – potentially in different treatment or diagnosis groups, it is also difficult for current computer technology to generate meaningful analysis that goes beyond measurement statistics and other calculated values.

[0004] While current computer technology cannot autonomously provide the full analysis, it is also difficult to integrate human input into the analysis. There is a vast amount of information available that current computer technologies cannot integrate efficiently since this would require a great deal of manual intervention by an experienced and skilled analyst, and will not be comprehensive.

[0005] Therefore, there is a need for an improved computer system that can provide meaningful analysis of measurement data for multiple samples while using input from a user to guide that analysis.

Summary

[0006] This disclosure provides methods for data analysis which are guided by user input. These methods are based on user input that identifies a subset of datapoints and generates a prompt to a machine learning (ML) model, trained to generate natural language text, including the measurement data of the subset of the multiple datapoints as selected by the user. The ML model is trained on a large corpus of texts in applicable technical fields and can therefore produce a meaningful analysis that incorporates available publications and other input. This way, the disclosed methods leverage the capabilities of trained ML models to provide an analysis of data, as selected by the user, in the form of natural language output text.

[0007] A method for generating a natural language output text characterising one or more biological samples comprises:

receiving user input from a user indicative of a selected subset of datapoints relating to the one or more biological samples, the subset of datapoints being selected on a user interface that graphically presents the datapoints to the user;

creating a natural language prompt for a machine learning model trained to generate natural language text, the natural language prompt comprising measurement data related to the selected subset of the datapoints; and

evaluating the machine learning model on the natural language prompt to generate the natural language output text characterising the one or more biological samples.

[0008] In some embodiments, the measurement data comprises quantitative data or qualitative data or both.

[0009] In some embodiments, the datapoints represent samples from different individuals.

[0010] In some embodiments, the prompt further comprises input data related to the one or more biological samples that is independent from the measurement data.

[0011] In some embodiments, the prompt further comprises input data provided by a user through the user interface.

[0012] In some embodiments, the method further comprises creating a further natural language prompt for the machine learning model, the further natural language prompt comprises information about an experiment that generates the measurements and text that causes the machine learning model to generate a further natural language output text characterising the experiment.

[0013] In some embodiments, the method further comprises creating a graphical user interface, the graphical user interface comprising multiple graphical data elements, each of the multiple graphical data elements representing one of the datapoints of the measurement data, the multiple graphical data elements being arranged in two dimensions on the graphical user interface.

[0014] In some embodiments, the user input is indicative of an area on the graphical user interface selected by the user and the method further comprises selecting the subset of the datapoints represented by the multiple graphical data elements that are within the area selected by the user.

[0015] In some embodiments, each data point represents one of multiple genes; the measurement data comprises sequencing data; and the prompt comprises a name for each of the multiple genes.

[0016] In some embodiments, the sequencing data comprises expression data of the multiple genes, the graphical data elements are visually formatted to represent the expression data, and the natural language prompt text comprises number values indicating the expression data for the subset of genes.

[0017] In some embodiments, each of the graphical data elements represents a step of a biological pathway.

[0018] In some embodiments, the method further comprises identifying one or more biological pathways that are related to a change in expressions levels of the multiple genes as

indicated by the sequencing data; and arranging the graphical data elements to represent the one or more biological pathways.

[0019] In some embodiments, each of the graphical data elements represents measurement data, including gene expression data, from a respective cell from one sample from an individual.

[0020] In some embodiments, the method further comprises performing a method of data analysis on the measurement data to arrange the graphical data elements in the two dimensions.

[0021] In some embodiments, the method further comprises calculating a position of each of the graphical data elements based on an output of the method of data analysis.

[0022] In some embodiments, the graphical data elements are points of a scatter plot that are arranged in two dimensions.

[0023] In some embodiments, the datapoints in the subset are selected by the user by drawing a free-form shape that encompasses the subset.

[0024] In some embodiments, the method further comprises calculating a score for each datapoint of the sub-set of datapoints.

[0025] In some embodiments, the method further comprises filtering the sub-set of datapoints based on the score to reduce a number of datapoints that is provided as the prompt to the machine learning model.

[0026] In some embodiments, the method further comprises ordering the measurement data in the prompt based on the score of the sub-set of multiple datapoints.

[0027] In some embodiments, the measurement data is obtained from an experiment of sampling a population of individuals with a sequencer and the prompt comprises text obtained from a gene database that is independent from the experiment.

[0028] A computer system for generating a natural language output text characterising one or more biological samples comprises one or more processors configured to perform the steps of:

receiving user input from a user indicative of a selected subset of datapoints relating to the one or more biological samples, the subset of datapoints being selected on a user interface that graphically presents the datapoints to the user;

creating a natural language prompt for a machine learning model trained to generate natural language text, the natural language prompt comprising measurement data related to the selected subset of the datapoints; and

evaluating the machine learning model on the natural language prompt to generate the natural language output text characterising the one or more biological samples.

[0029] A method for providing an interactive user interface with measurement data relating to one or more biological samples comprises:

creating a graphical user interface, the graphical user interface comprising multiple graphical data elements, each of the multiple graphical data elements representing a datapoint of the biological data, the multiple graphical data elements being arranged in two dimensions on the graphical user interface;

receiving user input from the user indicative of an area on the graphical user interface selected by the user;

selecting a subset of the datapoints represented by the multiple graphical data elements that are within the area selected by the user;

providing the measurement data related to the subset of datapoints within the area selected by the user as a prompt to a machine learning model trained to generate natural language text;

evaluating the machine learning model on the prompt to generate an output text that summarises the subset; and

presenting the output text to the user on the graphical user interface.

[0030] Optional features provided in relation to the method are equally optional features in relation to other embodiments, such as the computer system and other methods.

Brief Description of Drawings

[0031] Figure 1 illustrates a biological pathway with user selection of genes.

[0032] Figure 2 illustrates a computer system with processing modules.

[0033] Figure 3 illustrates a method for characterising sequencing data.

[0034] Figure 4 illustrates a computer system for analysing sequencing data.

[0035] Figure 5 illustrates a volcano plot with a user-selected subset of datapoints.

[0036] Figure 6 illustrates graphical data elements arranged according to the result of a principle component analysis and a user-selected subset of datapoints.

Description of Embodiments

[0037] As stated above, existing computer systems are not well adapted to providing a descriptive analysis of biological measurement data, such as sequencing data, because the number of data sources is very high and heterogeneous and the amount of data is immense. At the same time, human users are overwhelmed by the amount of data and also struggle to derive meaningful analyses. This is especially the case where multiple genes are involved, for example.

[0038] At the same time, there has been an emergence of trained generative language models that can be trained on a large amount of text data. However, those trained language models are not readily usable for the analysis of biological measurement data because such data is comprised of measurements, quantitative or qualitative, such as individual symbols (e.g., GATC), numerical values or strings of characters, and not natural language. Therefore, the measurement data cannot be interpreted readily by a trained language model.

[0039] To address these difficulties, the present disclosure provides methods where a user selects a subset of datapoints, such as genes, samples or experimentally-determined quantities, measurements or observations, and the computer system generates a natural language prompt including the subset and a representation of the measurement data, such as gene expression

data or titre data, in a form that can be interpreted by the trained language model, which is a ML model trained to generate text. In other words, the raw data is partially analysed to get it into a form that the trained language model can understand. Now, the trained language model can ingest the natural language prompt and draw on the training on medical literature to generate an output that, in effect, analyses the measurement data.

[0040] Figure 1 illustrates a biological pathway 100 as may be displayed on a user interface to a user. Each rectangle in the pathway 100 denotes one datapoint that represents one protein that is linked to a gene in the human genome, noting that this disclosure equally applies to any other species. The rectangles are filled with diagonal hatching to indicate an upregulated gene and a square hatching to indicate a downregulated gene. The user can now draw a freeform area 101 to select a subset of genes. The computer system then constructs a natural language prompt to be used as an input prompt to a trained ML model. That model then provides as its output an analysis of the sequencing data in natural language. As a result, the output is readily understandable by the user, which may be a researcher, clinician or other user. Further, the output can incorporate a vast amount of medical literature on which the model has been trained. It is noted that the output may be provided to the user as a stream that appears as the ML model generates the output or as a complete output once the ML generation is complete. It the example of the stream output, it may be possible for the user to stop the generation during generation and before the output generation has finished. Further, the user interface may comprise an input text box where a user can enter follow-up questions or other text prompts that are provided to the ML model in the context (the same chat) of the selected datapoints.

[0041] Figure 2 illustrates a computer system 200 comprising a bioinformatics pipeline module 201, a user interface module 202, a prompt generation module 203 and a language model module 204. Each of the modules of computer system 200 may be implemented as software modules, such as classes, functions, libraries, etc. as well as separate services or servers, such as remote or distributed services or servers providing application programming interfaces (API). As such, the bioinformatics pipeline module 201 receives the measurement data from laboratory equipment, such as the sequencer that may run an RNA sequencing process and, for example, a sequencing by synthesis process for RNA or a mass spectrometry process for sequencing proteins. The bioinformatics pipeline 201 may perform the mapping of short reads to a reference genome or the assembly of amino acid fragments into proteins.

Further, the bioinformatics pipeline 201 may determine variants or may calculate quantitative data, such as expression information, from the processed sequencing data.

[0042] The user interface module 202 receives the processed measurement data from the bioinformatics pipeline module 201 and creates a user interface that enables the user to select a subset of datapoints, such as genes. Such a selection should be based on the measurement data and therefore, it is advantageous to present the measurement data to the user in a suitable way that enables the user to select the subset of datapoints directly from the representation of the measurement data.

[0043] In one example, the sequencing data is presented in the form of a biological pathway. The user may be able select one or more pathways out of all known pathways. In other examples, the computer system 200 selects one or more pathways that are predominantly affected by the observed changes in the sequencing data compared to a baseline. For example, with RNA expression levels, the computer system 200 can select one or more pathways that are significantly affected by up- or downregulated genes.

[0044] The user may then be presented with a list of affected pathways to select from. In other examples, affected pathways are shown to the user as illustrated in Figure 1. The pathways are formatted such that the user can see the relevant aspects in the sequencing data, such as by shading or colour-coding genes in the pathway to show up- or downregulations. The user can then click on genes, or draw a bounding region around genes that are of particular interest. Those genes may be mainly those that are affected by up- or downregulation, or may be genes that are not regulated but otherwise appear possibly interesting to the user. In this way, the user can select the subset of genes in the same user interface that also presents the sequencing information, which is more useful than separate user interfaces or dropdown lists, tables, etc. More particularly, the user interface reduces both time and mental energy required to manage the transitions between different stages of the analysis, thereby making the analysis more intuitive and far more time can be directed to the important part of the problem which is interpreting and understanding the outcome.

[0045] In other examples, the user interface indicates where variants have been detected in DNA sequencing data and the shading or colour of the rectangles may indicate variants. This

way, the user can select those genes with variants, such as single nucleotide polymorphisms (SNP), structural variants, copy number variations, etc.

[0046] The user interface module 202 receives the selection from the user. This selection may be an array of selected datapoints representing genes or may be an array of coordinates that define an area in the user interface. In this way, it is advantageously possible to have a separate software program generate the graphical visualisation of the pathway as an image object (e.g., jpg or png) and then the user interface module 202 creates an overlay of invisible elements with calculated coordinates so that they overlay the corresponding genes in the pathway image object. In response to receiving coordinate data indicating a selection area, the user interface module 202 then calculates which of the invisible elements are within the area and creates the subset of genes this way. In other examples, the user interface module 202 receives an image area and retrieves from an image the image data that is in that area, and analysis the image directly without the need for the invisible layer. For example, the user interface module 202 may perform optical character recognition to extract identifiers of the selected datapoints (such as gene names) from the image data.

[0047] It is noted that any number of datapoints may be selected by the user, which includes one, more than one or all datapoints. Selecting all datapoints may be achieved explicitly by drawing a freeform shape around all datapoints or by activating a button that selects all datapoints for the user, or implicitly by selecting no datapoint and activating a button that starts the interpretation process and the user interface will automatically indicate that all datapoints are selected.

[0048] The user interface module 202 then passes the subset of datapoints representing genes to the prompt generation module 203, which creates a natural language prompt. That natural language prompt includes the names of the genes and corresponding measurement data, such as expression values of each gene, as number values in the natural language prompt. The advantage of the natural language prompt is that it is digestible by the language model module 204 that has been trained on other natural language content, such as medical publications. As a result, the language model module 204, evaluated on the prompt, generates a natural language output that analyses the measurement data. That output is now readily understandable by the user and requires no further numerical analysis or calculations. Instead,

the output provides observations, conclusions and interpretations of the sequencing data as well as a set of possible follow-up questions based on the output.

[0049] Figure 3 illustrates a method 300 for generating a natural language output text characterising measurement data relating to multiple biological samples, such as individuals. As discussed above, the measurement data can be any numerical or qualitative sequencing data that provides information about the sequencing of biological material for each of multiple genes. This may include DNA variants, RNA expression profiles, protein abundances, methylation and others.

[0050] In other examples, each datapoint represents a sample from a different individual. In that example, the measurement data may comprise any qualitative or quantitative measurement data. This may include a qualitative indicator of the presence of a particular characteristic, a quantitative titre value, an indication that a concentration of a particular compound or protein is above a predefined threshold, different structures of a protein, presence of single nucleotide polymorphisms (SNPs) or other genetic variants or biological markers, and other measurable quantity or quality. This also includes qualitative or quantitative polymerase chain reaction (PCR) analysis measurements.

[0051] In yet another example, the measurement data is single-cell RNA sequencing data. This is useful in cancer-related activities where each cell may have a different expression profile. This way, the measurement data includes RNA expression data for multiple cells. The user interface then provides a graphical representation of the many cells in a way that allows the user to select a sub-set of cells for further analysis. That is, each data point represents one single cell in e.g. a patient blood sample. The data user interface module 202 may perform a dimensional reduction (similar to PCA) to arrange the cells in two-dimensions, and also perform clustering analysis to identify groups of cells with similar gene expression profiles. Individual clusters may be colour coded (and potentially labelled), based on further analysis of the underlying data. The user then selects clusters or subclusters, or sets of clusters, which are then further interrogated with the help of generative AI as described above. Figure 7 illustrates an example for single cell RNA-Seq data. This is a UMAP, which is similar to principle components analysis (PCA). Each dot is a single cell and contains an entire transcriptome worth of information. Clusters of cells with similar expression profiles are interpreted as cell types. The disclosed method can extract the underlying expression data per

cluster to perform differential expression analysis for specific cell types and subsets. The user can then select cell types or clusters.

[0052] Figure 8 illustrates a further chart of single cell RNA-Seq data. This chart is called a dot plot. The clusters/cell types are drawn on one axis and a set of variable genes on the other. Gene expression is indicated by size and colour of the dots. The proposed methods can use the expression data to determine cell type/cluster identity.

[0053] The method 300 may be performed by a processor executing program code that causes the processor to perform method 300. As such, the processor receives user input from a user indicative of a selected subset of datapoints representing genes, samples, etc. As stated above, the user input may be an array or list of datapoints, or an indication of an area or coordinates that the processor can use to determine datapoints that are located within that area in the user interface. The user interface may be generated on a local machine or on a browser application of a remote client. In the latter case, "generating the user interface" comprises creating hypertext markup language (HTML) code that can be transmitted to the client and interpreted by the browser application. Equally, the browser application can monitor user interaction and send data back to the server that represents the user interaction with the user interface. As such, the user input may be in the form of a GET or POST routine or other call.

[0054] The generation of the user interface may be preceded by executing a bioinformatics pipeline, such as a pipeline that includes tools like TopHat and Cufflinks or DESeq2 (for differential expression analysis, i.e. quantitative statistical comparison between two groups of samples). These are free, open-source software tools for gene discovery and comprehensive expression analysis of high-throughput mRNA sequencing (RNA-seq) data. Together, they allow biologists to identify new genes and new splice variants of known ones, as well as compare gene and transcript expression under two or more conditions. Such tools may store their analysis data, such as differential expression data, in a file on a file system or web storage, such as Amazon Web Services (AWS) S3 using AWS Lambda for processing.

[0055] The computer system 200 may interface with these storage systems to retrieve sequencing data for the selected genes. In some cases, the sequencing data is numerical data in a machine-readable form without substantial natural language text. That is, the data may comprise multiple records and each record comprises the gene name as a label and the

differential expression value and potentially other data values in the record of that gene. Computer system 200 queries the data service for the list of genes and obtains a set of records that correspond to the selected genes.

[0056] In addition to the sequencing data, computer system 200 may also connect with other knowledge bases related to the selected genes. For example, computer system 200 may obtain natural language text that is created for reading by human users as opposed to the sequencing data that is created for other computer systems for further processing. The natural language text for each gene may comprise medical publication text or other expert knowledge expressed in natural language, such as a verbal description of each selected gene or a verbal summary of medical relevance. For example, the natural language text may be from a gene database that is independent from the experiment. This means the gene database has been created previously or independently from the experiment so that the sequencing data does not directly influence the content of the gene database.

[0057] For other measurement data, the computer system 200 may perform other pre-processing to display the measurement data on a user interface. For example, where the datapoints represent samples, the computer system 200 may create a two-dimensional arrangement, where each datapoint is a dot. This may be a scatter plot where the datapoints are arranged along two axes according to quantitative measurements of two parameters. In other examples, the plot is a volcano plot. Figure 5 illustrates a volcano plot with datapoints shows as dots and a user-selected area 501 that defines the subset of datapoints.

[0058] In yet another example, computer system 200 performs a dimension reduction procedure, such as principle component analysis (PCA), to reduce the number of dimensions to two from a potentially large number of measured parameters. Figure 6 illustrates a data plot of datapoints after PCA, where the user has defined an area 601 to create a subset of datapoints. While the datapoints are arranged according to their principal components (i.e. linear combination of originally measured parameters), each individual datapoint still relates to one of the original samples. Therefore, by selecting a subset of datapoints on the PCA space spanned by the first two principle components, the user selects a subset of the original samples. Therefore, computer system 200 can include the measurement data of those samples of the selected subset into the prompt. The user interface further comprises control elements

that enable the user to configure how to colour the graphical elements and how to cluster them.

[0059] In some examples (like in Figure 5 and 6), the datapoints in the graphical user element, i.e. the graphical elements, are coloured or otherwise visually modified to represent labels to the user. These labels may indicate groups, such as individuals that have received a respective treatment or groups that have other common characteristics, such as traits, phenotypes, genotypes, diagnosis, treatments etc.

[0060] For measurement data other than sequencing data and datapoints other than genes, computer system 200 can also retrieve information about the individuals that has been recorded but is not a "measurement" as such, e.g. age, sex, etc. This may also be referred to as metadata (for example via verbal questions, observations, or filling out a form), as opposed to data which is "measured" by a device such as an NGS apparatus. Metadata may further include race, diagnoses, genotype, phenotype, as well as literature information and others.

[0061] Based on the retrieved information, computer system 200 creates 302 a natural language prompt for a trained ML model. The ML model may be a large language model, such as a generative pre-trained transformer (GPT) including GPT-3 or GPT-4. The ML model is trained on a large corpus of text including most publicly accessible documents. This way, the ML model is trained to ingest and create most different types of texts. One example ML model is ChatGPT provided by OpenAI. Other examples include Microsoft's Bing or ChatSonic by Writesonic, Inc.

[0062] The input prompt to the generative natural language model is a natural language text that can be composed freely to provide instructions to the generative natural language model. For example, the prompt may be "explain the role of monocyte chemoattractant protein 1 in rheumatoid arthritis". It is noted that the output of the generative language model depends significantly on the input prompt. Therefore, this disclosure provides a method for generating a prompt that causes a trained generative language model to provide a meaningful analysis of measurement data that would otherwise be very difficult to interpret by a human clinician.

[0063] According to the present disclosure, the natural language prompt may comprise datapoint labels, such as names of the subset of the multiple genes, such as monocyte

chemoattractant protein 1. This prompts the generative language model to create an output related to that label. The prompt further comprises the measurement data, which includes the numerical or quantitative values from the measurement equipment or sequencing pipeline. This prompts the generative language model to create an output considering the actual measurement data (e.g., sequencing data) for the individual, or any other kind of sample that is not recognisably from an individual human or animal - e.g. cell culture, bacteria, plant, environmental swab, etc. from which the measurement data has been obtained.

[0064] Further, the prompt may comprise natural language prompt text for each of the subset of the multiple genes. The natural language prompt text may comprise natural language instructions and/or connections between the measurement data of the different datapoints (e.g., different gene expressions or titres) so that the prompt is ingestible and grammatically sufficiently correct for the trained generative language model. The prompt text may vary for different trained generative language models. An example prompt that has been found to work well with ChatGPT is:

```
Consider an RNA-Seq experiment.
In a minute I will give you data, and your instruction will be to interpret
the experiment by explaining the biological impact and implications of
condition 2 ({condition2}) in relation to condition 1 ({condition1}) on a
specific biological pathway: {pathway} ({map_id})
Description of pathway:
{pathway_description}
I am about to give you a list of genes. please write out this list and
mention that these are the genes of interest. I only want you to focus on
this particular subset of genes. Here is the list: {selected_gene_list}
The pathway enrichment stats are:
{stats}
Some metadata might be provided here (but if not, don't worry about it):
{metadata}
The experimental variables are summarised in this table:
{variables}
The samples are summarised in this table:
{samples}
I am about to give you the most important data, describing differential
gene expression between the two experimental conditions. Note: A positive
value for log2foldchange here means that a gene was up-regulated in
condition 2 (in other words it was expressed at a higher level in condition
```

```
2 than in condition 1). Conversely, a negative value for log2foldchange
here means that a gene was down-regulated in condition 2. Here is that
data:{gene_data}
In the differential gene expression table, the column "Score" is an attempt
to measure the importance of each gene, where bigger is better and should
be used in combination with the other info. don't specifically mention this
"score", but you may use it to help prioritise things. Ignore "nan" genes
(genes without proper names).
Stay focused on this pathway and the specific genes of interest and their
known functions and interactions. You are writing for a sophisticated
audience of biological scientists. Don't just summarise the metadata, focus
on the meaning of the results. You may also include some suggestions for
further research. Rather than dwelling on generic responses, you should
focus on interpretations that refer to the experimental results.
Next, answer the following question: What additional information can I
provide to make this analysis even more insightful? Keep in mind, we can
only add a couple of hundred more words, so it will be to explain things
like the motivation of the experiment or providing an explanation for any
components you do not fully understand, etc. Include a main title and
subtitles for each paragraph or section. Use bold text (markdown) for
titles and gene names. With markdown, use H1 for the main title and H3 for
subtitles. Do not include citations or hyperlinks.
```

[0065] Once the prompt is constructed, the computer system 200 evaluates 303 the trained generative language model on the natural language prompt to generate the natural language output text characterising the biological samples.

[0066] As set out above, a number of representations may be used to enable user selection of datapoints (e.g., genes, measurements, plot elements). For example, it is possible to perform data analysis on the sequencing data to arrange the graphical data elements in the two dimensions. Some of these methods, calculate a position of each of the graphical data elements based on an output of the method of data analysis. For example, the graphical data elements may be points of a scatter plot that are arranged in two dimensions and the arrangement in the two dimensions is based on the sequencing data so that the position represents the numerical value in x and y direction. In one example, this representation is a volcano plot, which is a type of scatterplot that shows statistical significance (P value) versus magnitude of change (fold change). The user can then draw an area or otherwise select points (representing genes) that have the desired p-value and fold change, such as a large p-value

and a large fold change. The points representing the genes may further be colour coded to further guide the user in the selection.

[0067] Further, the disclosed method may be used in the area of spatial transcriptomics. In that example, the selection of datapoints is referred to as "spots" (as opposed to "cells" as used in single cell sequencing). A "spot" is a small region of tissue, but not necessarily a single cell. It may comprise a few different cells (or even parts of cells), and the expression pattern of a spot is the sum or average of the cells it consists of. Further, the selection may be on an image that is a composite of a scatterplot and a microscope image (example shown in Fig.). Perhaps the term "datapoints" may suffice, but it might also need to be mentioned that the user may be selecting regions or structures identified in the image, as opposed to purely selecting "datapoints". 3) future versions are likely to include additional image processing methods such as automatic detection of features (image segmentation), and a user may select one or more "segments" to analyse (segments also contain sets of datapoints). In this way it is quite similar to selection of clusters. Other object terminology that might be relevant to user interaction: structures, regions, areas of interest, elements, features.

[0068] Figure 9 illustrates spatial RNA-Seq data. Similar to single cell, but now each dot is a "spot" and may consist of a few cells or even parts of cells (depending on the size and location of the spot). The analysis is similar as for single cell (i.e. the method starts with UMAP and dot plot), except now every spot also has coordinates that map it to a location on a microscope image. In Figure 9, the chart on the top is for a single gene and uses a quantitative colour palette (shown in black and white in Figure 9 for reproducibility). The one on the right uses a qualitative colour palette (again shown in black and white) to show the different clusters (interpreted as different cell types).

[0069] While examples above relate to datapoints displayed as graphs, it is noted that objects other than graphs that equally be used for selection. The user can also interact with these object and the AI can interpret them. Examples include tables, flow charts, blocks of text, images, forms (e.g. a collection of user input such as dropdowns, check buttons, text boxes etc). In multiple instances the method calls on the AI to interpret purely table data (e.g. gene set enrichment results, differential expression results, Venn diagram segments), and these tables can also be interactive, thereby modifying the prompt (e.g. selection of one or more cells, rows or columns). An example where it occurs using only form data is the

experiment page, where the AI looks at the proposed experimental design and provides feedback. The elements in this case comprise variables, replicates, metadata and text descriptions, while interaction means filling out the form.

[0070] In yet a further example, the accuracy of the output may decline with a too large number of input datapoints, such as genes. Therefore, computer system 200 may calculate a score for each datapoint of the sub-set of the datapoints and filter the sub-set of the multiple datapoints based on the score to reduce a number of datapoints that is provided as the prompt to the generative language model.

[0071] In further examples, the output generated by the trained generative language model depends significantly on the order of the datapoints provided in the input prompt. Therefore, computer system 200 orders the measurement data in the prompt based on the score of the sub-set of datapoints.

[0072] In another example, computer system 200 creates a further natural language prompt for the machine learning model. That further natural language prompt comprises information about an experiment that generates the measurements and text that causes the machine learning model to generate a further natural language output text characterising the experiment. In this sense, the information about the experiment may comprise a description that may be a free, natural language text about knockout of cell lines, NGS tests, description of aims of the experiments and comparisons with hypotheses. This may further include reasons for which particular genes were chosen. It may further include details about the treatment of samples, such as samples transfected with different compounds and description about the use of the measurement data, such as the development of a vaccine against a specified disease.

[0073] The experimental data may be provided in a user interface that comprises an input box for the description mentioned above as well as input fields for data variables. Such as variables for treatment where the user can enter different treatment labels or knockout gene names as text values. The user interface may also comprise input fields for title, data source, data type (e.g., mRNA) and species. The user interface may also comprise input fields for entering a number of replicates for different groups.

[0074] The user interface may also comprise selection lists to form pairs of groups for differential expression analysis as well as selection of (potentially ranked) biological pathways).

[0075] The machine learning model is understood to be a model, such as a mathematical model, that receives input text (as referred to as 'prompts') or other inputs such as, but not limited to, image data and audio data and generates an output based on the input. The machine learning model may be of an architecture, such as, but not limited to, a neural network, for example. In general, machine learning models are 'trained' to learn and recognise patterns in an input and provide an output that is a prediction based on the training it has undergone. Training involves updating weights or parameters (as referred to as hyperparameters) of the machine learning model, which define the machine learning model, to minimise a loss value, thereby creating a trained machine learning model (in other words, a machine learning model trained to generate an output). This may involve a gradient descent and backpropagation method.

[0076] The machine learning model may be stored on data memory or a server by storing the weights that define the model. As such, the machine learning model may be referred to as a "memory model", given that it is defined by parameters (i.e., the weights) which can be stored on computer memory. In some embodiments, the machine learning model may be programmed on an integrated circuit, such as a field-programmable gate array (FPGA) or a graphics processing unit (GPU) such as an NVIDIA processing unit. In such an embodiment, processor 102 may not retrieve the parameters from the data memory. Instead, an input may be communicated from the processor to the integrated circuit and the integrated circuit may apply the machine learning model to the input and generate an output, which is then communicated to processor.

[0077] Integrated circuits, such as FPGAs and GPUs, can be used where flexibility, speed, and parallel processing capabilities are desired. In such an embodiment, the integrated circuit may be part of the system and may be considered as a "processor" or "processing unit", similar to a local CPU. Other implementations, such as application specific integrated circuits (ASIC) or neuromorphic architectures are equally useable.

[0078] There may be a number of different ways to invoke the machine learning model, which all fall within the meaning of "evaluating the machine learning model". For example, evaluating the model may involve calling an API routine to send the prompt to a server and the server then performs the calculations according to the trained machine learning model and returns the results. In other example, evaluating may involve issuing a command to local hardware, such as a local chip, device, machine learning accelerator (e.g., a USB device design to efficiently perform machine learning tasks or Nvidia's Deep Learning Accelerator (DLA)), etc., that has the trained machine learning model stored thereon and provides a command interface to interact with the model. It is also possible to have a local copy of the machine learning model available so that the calculations are performed by the main processor of the local machine. Other local, remote or distributed implementations are equally useable.

[0079] In some embodiments, the machine learning model may be trained on a broad range of different data such that it can be applied across a wide range of use cases. Such a machine learning model may be referred to as a "foundational model". Some foundational models that are applicable to the disclosed method include those that are publicly available and/or trained on publicly available data. In other embodiments, the machine learning model may be trained on a specific set of training data, in order to focus the generated outputs of the machine learning model to a specific task or area of interest. In yet other examples, the foundational model is further trained on the specific set of training data to improve the model in the area of interest.

[0080] In some embodiments, the machine learning model is a multimodal machine learning model, in which multiple inputs of different modalities (e.g., text, image data and audio data) are used to provide one or more generated outputs. An example of a multimodal machine learning model is an object detection model, which detects the location of a specific object (specified by input text, for example) in an image. This example model may generate output text that describes the location of the specified object in the image. Although the multimodal machine learning model can be evaluated on multiple input of different modalities, the multimodal machine learning model can also be evaluated on a single input and still generate an output based on the single input.

[0081] The machine learning model may be trained to generate output text based on input text and hence, may be a chat-based machine learning model. The machine learning model may also be referred to as a trained generative language model. Both the input and output text may be in the form of "natural language" (i.e., any language that occurs naturally in a human community by a process of use, such as spoken English, for example). Such machine learning models may be referred to as "chatbots". A chatbot (which may also be referred to as a chatterbot) is designed to mimic human conversation (using natural language) through text or voice interactions. More particular, the chatbot response to input natural language using output natural language. Examples of such chatbots currently include ChatGPT (using GPT-3 or GPT-4), Microsoft's Bing Chat/Copilot (which may use OpenAI's GPT-4) and Google's Bard.

[0082] In more specific examples, the machine learning model may be an LLM, may be based on an LLM, may include an LLM or may be derived from an LLM. An LLM is a type of artificial intelligence system, characterised by its massive training data and high volume hyperparameters. These language models ingest input text sourced from various sources and use fine-tuning to predict potential tokens or words. This enables them to perform various natural language processing tasks, including sentiment analysis, document classification, and lexical analysis, among others. However, their capabilities have extended beyond these tasks to encompass a broader range of applications and industries, such as chatbots, content generation, and even scientific research, demonstrating their versatility and growing significance in the field of AI.

[0083] LLMs like GPT (i.e., GPT-3, GPT-3.5, GPT-4, GPT-4o, GPT-4o-mini, DeepSeek) and LLaMA are built upon the transformer architecture. These models are pre-trained on vast volumes of text data collected from publicly available online resources, and have been specifically trained to understand and process natural language. However, they differ from traditional natural language processing (NLP) systems in how they handle NLP processing tasks. Instead of relying solely on pre-training and fine-tuning, LLMs excel in NLP tasks through the technique of prompt engineering. In prompt engineering, tasks are conveyed as text descriptions, and these descriptions are presented to the model for interpretation and the generation of corresponding responses.

[0084] The LLM may be an artificial neural network, such as transformer model (e.g., a generative pre-trained transformer) which utilises encoder and decoder networks. These LLMs are (pre-) trained using self-supervised learning and semi-supervised learning. In essence, LLMs are trained to predict what word comes next in a sequence of words which can be based on the semantic closeness of the words. LLMs form these predictions by 'tokenising' the words in the output text (i.e., converting the words into a vector of numbers). These tokens can also incorporate other information such as the position of the word in the input text and the information about the adjust words.

[0085] The process of creating these tokens is also referred to as 'encoding' or 'semantic encoding', as information regarding each word is essential encoded into a vector of numbers. The process of creating these token that incorporate the position of each word in the input text is referred to as "positioning embedding". The encoding process also includes embedding at least one learnable parameter, which is determined through training. The opposite process then occurs where the tokens are converted from a vector of numbers into words, which is referred to as 'decoding'. This gives the final output of the LLM in the form of text.

[0086] The machine learning model may also comprise an attention mechanism that applies weights to the tokens and may be characterized by its self-attention layers. These layers enable the model to assess the significance of words in an input relative to each other, thereby providing a more nuanced comprehension of the text. The attention mechanism may also include a scaled dot-product between different matrices generated by the model to calculate the weights. In simple terms, the attention mechanism allows a model to focus on different parts of the input when generating each element of the output. This dynamic focusing capability results in a more contextually aware model, producing better results in tasks like translation, summarization, or text generation. Other ways of achieving attention within the model would be equally possible.

[0087] As a result, the machine learning model can process a large number of input values, such as an input text paragraph, at one time rather than sequentially in order to consider the context of each word. Nevertheless, the overall number of parameters in the machine learning model is relatively large, which is the reason those models are referred to as large models, such as large language models or large action models. In some examples, a model is large if it

has more than 100 million parameters or more than 1 billion parameters or more than 1 trillion parameters.

[0088] In some embodiments, the machine learning model referred to in this disclosure may be multiple machine learning models that have a 'global' input and a 'global' output. In that sense, the multiple machine learning models may be "daisy chained" together, such that the output of one machine learning model becomes the input for the next machine learning model within the chain. Some of the multiple machine learning models may operate in parallel, rather than in a series or chain. Each of the multiple machine learning models may have its own memory and/or have access to a common memory that is shared become some or all of the multiple machine learning models. In this sense, the multiple machine learning models may resemble parallel processing or parallel computer, such a computer architecture with multiple CPU cores which can be operated in parallel.

[0089] Figure 4 illustrates a computer system 400 for analysing sequencing data, which relates to computer system 200 in Figure 2. As such, computer system 400 comprises a processor 401 connected to a non-volatile storage medium 402. The processor 401 may also comprise multiple processors 401 that are individually or in combination programmed to perform the disclosed methods. Non-volatile storage medium 402 comprises program storage 403 and data storage 403. Program storage 1403 stores computer code that, when executed, causes processor 1401 to perform the methods disclosed herein. In particular, computer code implements method 300 in Figure 3.

[0090] It will be appreciated by persons skilled in the art that numerous variations and/or modifications may be made to the above-described embodiments, without departing from the broad general scope of the present disclosure. The present embodiments are, therefore, to be considered in all respects as illustrative and not restrictive.

CLAIMS:

1. A method for generating a natural language output text characterising one or more biological samples , the method comprising:

receiving user input from a user indicative of a selected subset of datapoints relating to the one or more biological samples, the subset of datapoints being selected on a user interface that graphically presents the datapoints to the user;

creating a natural language prompt for a machine learning model trained to generate natural language text, the natural language prompt comprising measurement data related to the selected subset of the datapoints; and

evaluating the machine learning model on the natural language prompt to generate the natural language output text characterising the one or more biological samples.

2. The method of claim 1, wherein the measurement data comprises quantitative data or qualitative data or both.

3. The method of claim 1 or 2, wherein the datapoints represent samples from different individuals.

4. The method of any one of the preceding claims, wherein the prompt further comprises input data related to the one or more biological samples that is independent from the measurement data.

5. The method of any one of the preceding claims, wherein the prompt further comprises input data provided by a user through the user interface.

6. The method of any one of the preceding claims, wherein the method further comprises creating a further natural language prompt for the machine learning model, the further natural language prompt comprises information about an experiment that generates the measurements and text that causes the machine learning model to generate a further natural language output text characterising the experiment.

7. The method of any one of the preceding claims, wherein the method further comprises creating a graphical user interface, the graphical user interface comprising multiple graphical data elements, each of the multiple graphical data elements representing one of the

datapoints of the measurement data, the multiple graphical data elements being arranged in two dimensions on the graphical user interface.

8. The method of claim 7, wherein the user input is indicative of an area on the graphical user interface selected by the user and the method further comprises selecting the subset of the datapoints represented by the multiple graphical data elements that are within the area selected by the user.

9. The method of claim 7 or 8, wherein
each data point represents one of multiple genes;
the measurement data comprises sequencing data; and
the prompt comprises a name for each of the multiple genes.

10. The method of claim 9, wherein
the sequencing data comprises expression data of the multiple genes,
the graphical data elements are visually formatted to represent the expression data, and
the natural language prompt text comprises number values indicating the expression data for the subset of genes.

11. The method of any one of claims 7 to 10, wherein each of the graphical data elements represents a step of a biological pathway.

12. The method of claim 11, wherein the method further comprises:
identifying one or more biological pathways that are related to a change in expressions levels of the multiple genes as indicated by the sequencing data; and
arranging the graphical data elements to represent the one or more biological pathways.

13. The method of any one of claims 7 to 12, wherein each of the graphical data elements represents measurement data, including gene expression data, from a respective cell from one sample from an individual.

14. The method of any one of claims 7 to 13, wherein the method further comprises performing a method of data analysis on the measurement data to arrange the graphical data elements in the two dimensions.

15. The method of claim 14, wherein the method further comprises calculating a position of each of the graphical data elements based on an output of the method of data analysis.

16. The method of claim 14 or 15, wherein the graphical data elements are points of a scatter plot that are arranged in two dimensions.

17. The method of any one of the preceding claims, wherein the datapoints in the subset are selected by the user by drawing a free-form shape that encompasses the subset.

18. The method of any one of the preceding claims, wherein the method further comprises calculating a score for each datapoint of the sub-set of datapoints.

19. The method of claim 18, wherein the method further comprises filtering the sub-set of datapoints based on the score to reduce a number of datapoints that is provided as the prompt to the machine learning model.

20. The method of claim 19, wherein the method further comprises ordering the measurement data in the prompt based on the score of the sub-set of multiple datapoints.

21. The method of any one of the preceding claims, wherein the measurement data is obtained from an experiment of sampling a population of individuals with a sequencer and the prompt comprises text obtained from a gene database that is independent from the experiment.

22. A computer system for generating a natural language output text characterising one or more biological samples, the computer system comprising one or more processors configured to perform the steps of:

receiving user input from a user indicative of a selected subset of datapoints relating to the one or more biological samples, the subset of datapoints being selected on a user interface that graphically presents the datapoints to the user;

creating a natural language prompt for a machine learning model trained to generate natural language text, the natural language prompt comprising measurement data related to the selected subset of the datapoints; and

evaluating the machine learning model on the natural language prompt to generate the natural language output text characterising the one or more biological samples.

23. A method for providing an interactive user interface with measurement data relating to one or more biological samples, the method comprising:

creating a graphical user interface, the graphical user interface comprising multiple graphical data elements, each of the multiple graphical data elements representing a datapoint of the biological data, the multiple graphical data elements being arranged in two dimensions on the graphical user interface;

receiving user input from the user indicative of an area on the graphical user interface selected by the user;

selecting a subset of the datapoints represented by the multiple graphical data elements that are within the area selected by the user;

providing the measurement data related to the subset of datapoints within the area selected by the user as a prompt to a machine learning model trained to generate natural language text;

evaluating the machine learning model on the prompt to generate an output text that summarises the subset; and

presenting the output text to the user on the graphical user interface.