

Gazing into the Metaverse: Automated exploration and contextualization of metabolic data

This manuscript was automatically generated on March 10, 2020.

Authors

✉ Jordan A. Berg¹, ✉ Youjia Zhou², ✉ T. Cameron Waller^{1,3}, ✉ Yeyun Ouyang¹, ✉ Ian George¹, ✉ Tyler Van Ry^{1,4}, ✉ James Cox^{1,4}, ✉ Bei Wang², ✉ Jared Rutter^{1,5}

1. Department of Biochemistry, University of Utah
2. School of Computing, University of Utah
3. Division of Medical Genetics, Department of Medicine, School of Medicine, University of California San Diego
4. Metabolomics Core Facility, University of Utah
5. Howard Hughes Medical Institute, University of Utah

† To whom correspondence should be addressed:

* This is a draft of the eventual final manuscript, and should therefore be treated as such. Conclusions, along with author order and contributions, may change as the manuscript is finalized.

Abstract

Science has utilized a largely reductionist approach to understanding metabolic systems in the past. While such an approach was previously necessary due to technological limitations, current computer age technological advances paired with -omics experiments allow for the survey, modeling, and exploration the biological systems in detail. Yet, our ability to contextualize and extract the full extent of these enormous datasets continues to lag and often results in focusing on limited entities from a dataset. To address these challenges, we developed Metaboverse, a multi-omic computational analysis framework and application for the interactive exploration and automated extraction of potential regulatory events, patterns, and trends from user data within the context of the metabolic network. This framework will be foundational in increasing our ability to holistically understand static and temporal metabolic events and shifts and gene-metabolite intra-cooperativity, as well as ensure we obtain the maximum amount of information from our data. Metaboverse is freely available under a GPL-3.0 license at <https://github.com/Metaboverse/>.

Introduction

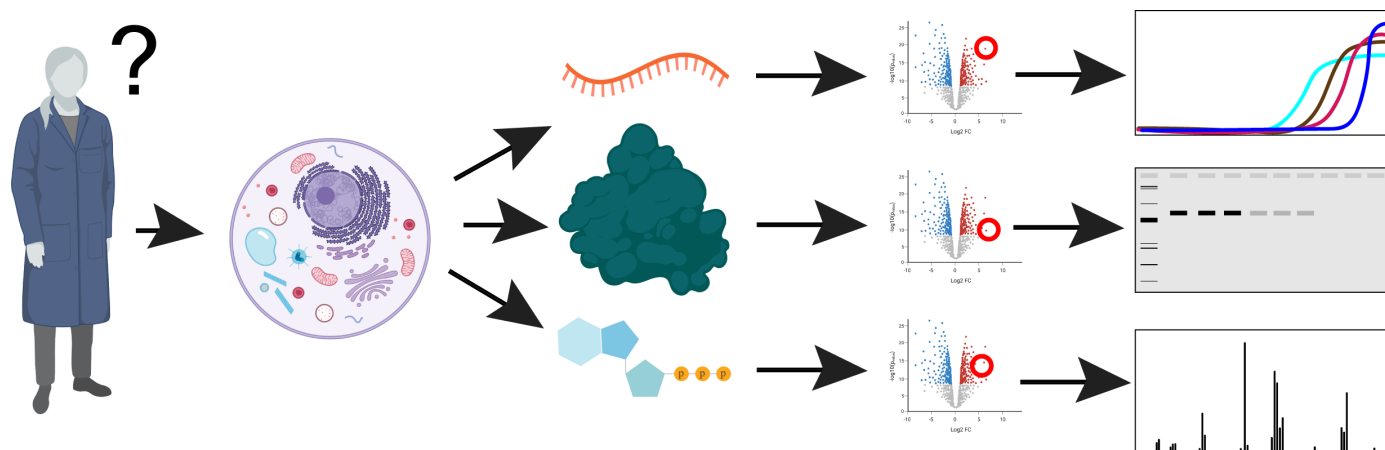
Metabolism is a complex network of reactions and interactions between genes, enzymes, complexes, and metabolites. To understand these complex components, scientists normally adopt a reductionist approach to teasing apart the characteristics and mechanics of these processes and how they fit into the larger picture of biology and disease. While a vital component in the scientific process, by doing so, many interesting properties of metabolism can be missed. For example, in differential gene expression analysis, researchers rely on thresholds of magnitude and statistical significance to prioritize genes for follow-up study. However, doing so can inadvertently limit the scope of study of metabolism when in fact metabolism is a highly interconnected system where distal components and their modulation can have rippling effects across the network. The current approach is analogous to telling the story of Little Red Riding Hood, but only by reading the 20 most frequent words used in the study. Certainly doing so efficiently highlights key words like “wolf” and “little red riding hood,” but also prevents a coherent story from being told and would make it difficult for someone who had never read the story of Little Red Riding Hood from comprehending the story.

Over the past decade, several computational tools have emerged and become popular for their focus on trying to solve these issues in data contextualization. We will highlight four, and while others exist, we focus on tools representative and most popular for their respective properties. First is MetaboAnalyst, which relies largely on set enrichment methods, or looking at the belongingness of sets of significantly changed analytes (i.e. metabolite, protein, or gene measurements), for extracting interesting information. While network visualization is available, its ability to extract regulatory information is limited, particularly in an automated fashion. Second is Cytoscape, which focuses on network representations of metabolism and other systems. While a variety of plug-ins are available for customizing analyses, again, pattern recognition and other features are lacking. MetExplore focuses on the curation of networks, and is particularly useful for collaborative annotation of emerging organisms. It additionally can layer experimental data on the network for visualization. Reactome, which Metaboverse uses for the curation of biological networks, also offers analytical tools for user data, but again relies on set enrichment or manual methods for identifying patterns. While all have their respective utility, there is still a need for tools that automate pattern and trend detection, especially when data is sparse, across metabolic networks in order to extract regulatory and other features from data.

In order to address these limitations in current conventions of metabolic data analysis, contextualization, and interpretation, we created the software application, *Metaboverse*, to aid users

in filling in the details of their model's metabolic story. *Metaboverse* is an interactive tool for exploratory data analysis that searches user data in the context of the metabolic network to identify interesting patterns and trends in the data. *Metaboverse* will aid scientists in formulating new hypotheses from their data and aid them in designing follow-up experiments for a deeper understanding of their model. *Metaboverse* operates across the entire metabolic network to quickly and automatically detect patterns and trends from a pre-designed pattern library, or can accept interactive input from the user where they can define certain patterns or trends they would like to identify across the global metabolic network. Figure 1 provides a graphical abstract to illustrate *Metaboverse's* role in the exploratory data analysis of biological data in the context of metabolism.

Traditional (reductionist)



Metaboverse (holistic)

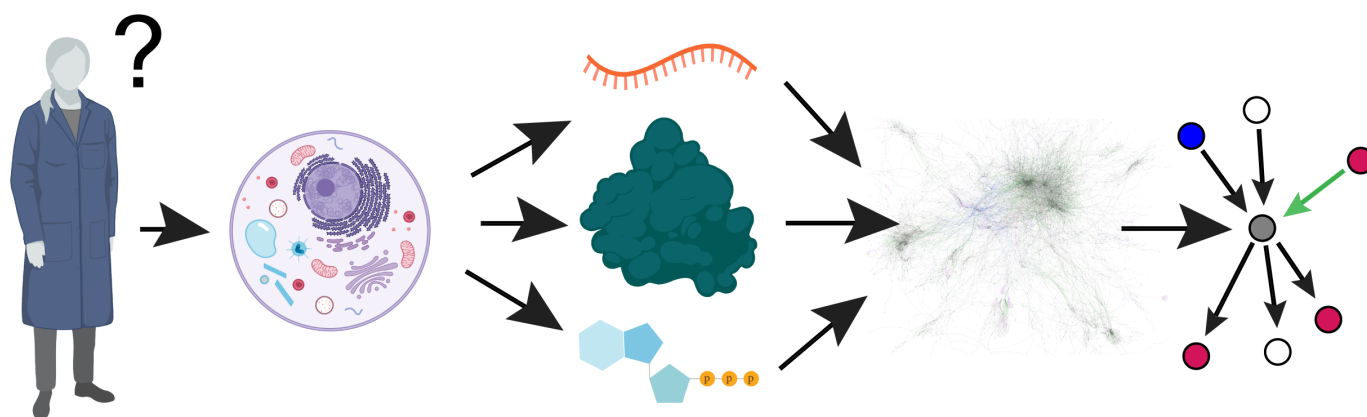


Figure 1: Metaboverse conceptual overview. Illustration comparing traditional metabolic data analysis methods and the holistic approaches that Metaboverse offers. Traditionally, when a scientist performs an -omics experiment, they tend to focus on a couple of features that are differentially regulated. Metaboverse inversely contextualizes the data across the metabolic network and identifies interesting regulatory patterns in the data.

In order to provide a platform for the exploration of single or multi-omic metabolic data, we developed several new computational features to aid in the aims discussed above. First, we developed a pattern search engine for the rapid and automated identification of patterns and trends in -omic data on the metabolic network. Conceptually, this search engine borrows principles of topological motif searching from graph theory. In the computational science context, a motif is simply a re-occurring pattern in network structure, or the organization of network entities and their relationships to one another. However, with -omic data, we are more interested in identifying patterns in

expression or abundance of genes, proteins, and metabolites. We therefore adapted this methodology to search the global metabolic patterns for interesting patterns in the network. For example, at a reaction the input may be high and the output low, indicating some sort of regulatory event occurring in the model. *Metaboverse* will search the global network from a pre-defined library of regulatory patterns and return an ordered graphical table of conserved patterns. Users will also be able to design their own patterns through an interactive pattern drawing tool, and even design specific scenarios that are cognizant of feature type. For example, one might be interested in a pattern where a protein displays higher expression, but the resulting metabolite is decreased. The user can also define multi-step patterns that may occur over two or more reactions.

Another feature introduced in *Metaboverse* allows for the interactive exploration of specific reactions or reaction entities with on-the-fly pattern search analysis. Users can explore specific pathways of interest and look for other interesting patterns and trends in pathways of interest. Users can also select a specific metabolite, protein, complex, or gene and explore patterns across pathways in a feature neighborhood search. For example, a user might identify a change in one metabolite in a particular pathway and want to explore what distal effects this change has in other pathways that use this metabolite. This functionality moves the analysis away from our traditional, strictly defined pathway approach to analysis, and helps contextualize the far-reaching effects changes in metabolism can have across the classical pathways in metabolism.

One challenge in metabolomics data analysis is sparsity of data points. While thousands of metabolites exist in human metabolism, the current state of the technology for determining which mass spectra belong to which metabolite can be challenging and often results in a limited number of data points being quantified. These can lead to gaps in the metabolic network which can be challenging to explore and analyze. We therefore introduce a reaction collapse feature that allows for summarization of reactions for which data is missing. This methodology can use RNA-seq data to inform prioritization of particular paths. For example, one metabolite may be converted to a downstream product in two different manners, but by using the gene expression data from a model, one can determine that one path is active while the other is inactive.

Metaboverse is designed to handle standard two-condition experiments, flux metabolomics, and time-course experiments. Time-course inputs can be single-omics, or static RNA-seq and/or proteomics with multiple metabolomic time-points. Users input fold change and statistical measures from their respective -omics, and *Metaboverse* reconciles the inputs for layering on the metabolic network. *Metaboverse* can handle data from a variety of model organisms, including humans, mouse, yeast, zebrafish, and more. The foundational curation of *Metaboverse* is built on the Reactome curations of metabolism, so any of the 90+ species available on that platform are also available within the *Metaboverse* environment. In order to validate these methodologies available in *Metaboverse* we analyzed two-condition, flux metabolomics, and time-course datasets and provide vignettes that highlight *Metaboverse's* reliability in extracting canonical features, as well as novel features and patterns, from well-defined biological models. We outline the technical specs for computational biologists in the methods section, and the biological utilities in the main text for wet bench biologists. We intend that *Metaboverse* will be foundational in our ability to more deeply and holistically explore metabolism and aid in our ability to provide more context within metabolic models.

Results

Metaboverse is a dynamic, user-friendly tool for the exploration of high-throughput biological data in organism-specific pathways.

Overview.

We designed *Metaboverse* as a light-weight, self-contained app for the dynamic exploration of high-throughput biological data. The pathway curations are derived from Reactome, and as of writing, is capable of analyzing data for ## species. A user begins by providing a previously curated Metaboverse file, or the desired output location for a new curation, and selecting their organism of interest. Next, the user provides the relevant gene, protein, and/or metabolite datasets they would like layered onto the global reaction network of their organism of interest. These data categories can be extended to any dataset that uses the relevant mapping IDs; for example, one could provide ribosome profiling translation efficiency values mapped to the appropriate gene IDs for layering onto the network and downstream analysis. During this step, the user will also specify a few experimental parameters for consideration during downstream analysis and visualization. Following these user inputs, the organism's network, data processing, and motif analysis (discussed further below) is curated and a curation file is output for future analysis.

Figure ##. Overview

Filling in the missing space.

Missing data points, particularly in metabolomics experiments, are frequent and can make analysis of pathways and identification of regulatory patterns in the network challenging. We therefore developed a reaction compression algorithm (detailed more in the Methods section) that collapses up to three reactions with missing data points if they can be bridged with known data on the distal ends of the reaction path. These reactions, or pseudo-reactions are visually distinct during visualization of the network and allow the user to quickly identify interesting patterns in the network, learn what that pseudo-reaction was summarizing, and generate additional hypotheses based on the available information and lack of information.

Rapid identification of interesting regulatory patterns in the reaction network.

Following network curation, the user can visualize available reaction motifs identified across the global reaction network. In *Metaboverse*, we define a motif as a regulatory pattern identified across a reaction or pseudo-reaction. *Metaboverse* contains a library of default motifs to search the network for, and users can define custom motifs they would like to identify across the global network. *Metaboverse* then displays these motifs within a "stamp view", where available motifs are ranked and displayed by magnitude. For a given pattern, the user can then explore each pathway this particular motif is found in. Motif analysis of the global regulatory network will allow users to rapidly identify interesting features in the data, particularly patterns between canonical pathways or in other pathways that may not be an initial focus in their research. In the data vignettes below, we demonstrate this utility further.

Dynamic visualization of organism-specific reaction pathways.

Following curation of the global network as described above, the user can manually search individual canonical pathways or individual entities and their reaction neighborhoods. For a given selection, all relevant reactions that are annotated as a part of that pathway will be graphed, along with their core

input (reactant) and output (product) components. In addition to these core elements, known catalysts and inhibitors are included, as well as the component proteins, genes, and metabolites known to form a functional complex as part of a reaction. Labels can be toggled on or off in the display, and the user can switch between viewing the values or statistics associated with each data point with their relevant color mapping. In cases where a gene value is known, but its protein value is unmeasured, the protein value will be inferred using aggregated gene component values. The same is then done for functional complexes using their inferred or known component values. Relevant pathway and analytical metadata is also displayed. Other information, such as identified motifs found in the pathway can also be found and expanded for further exploration in a new window. Aids for visualization are also available, such as the ability to remove nodes from visualization that contain a high number of relationships to other network features such as that these nodes, which act as hubs in the network, do not lead to cluttered representations of the network. Often, these hub nodes are ubiquitous features such as water and proton which may be of limited interest to the user during data visualization. Compartment domains are also graphed to include a relevant reactions and their components that occur in a given cellular compartment.

Visualization of downstream effects of network perturbations.

Users may be interested a particular metabolite or protein and the downstream effects its perturbation has on related pathways. By double-clicking a node of interest, or by selecting the entity name from the drop-down menu, the user can explore all downstream effects across all pathways in the global network. The user can also define how many neighborhoods to display such that one can visualize two or more reaction steps downstream of the selected entity.

Data vignettes

In order to demonstrate the added utility of *Metaboverse* to the community that is not currently available in other tools, we used *Metaboverse* to analyze a series of public and new datasets. From the vignettes provided below, we show that *Metaboverse* not only is able to identify points of interest previously described or expected, but can rapidly identify for the user unexpected and systematic regulatory patterns in a reaction network context.

1. Static (Ian)

Figure 3. Data
Supp Table 1. Motif results

2. Time-course (Yeyun)

Figure 4. Data
Supp Table 2. Motif results

3. Flux data (Cameron)

Figure 5. Data
Supp Table 3. Motif results

Performance

Table 1. Performance break-down
Table 2. Comparison to existing tools

Discussion

We hope that this tool will bring a new perspective to users' data and help draw the connections needed to aid them in extracting new and exciting hypotheses from their data that would be difficult to do without this tool.

Methods

A tutorial for how to use *Metaboverse* can be found at metaboverse.readthedocs.io/getting-started.

1. Network Curation

Biological networks are curated using the current version of the Reactome database. In particular, the pathway records for each species, complex component and interaction data, Ensembl, and UniProt Reactome mapping tables are integrated into the network database for *Metaboverse*. Additionally, the ChEBI database names table (ftp://ftp.ebi.ac.uk/pub/databases/chebi/Flat_file_tab_delimited/names.tsv.gz) is integrated. These data are used to generate a series of mapping dictionaries for entities to reactions and reactions to pathways for curation of the global network.

After the relevant information is parsed from each table or record, the global network is propagated using the NetworkX networking framework [cite:networkx] to generate nodes for each reaction and reaction component, and edges connecting components to the appropriate reactions. In some cases, a separate ID is used to generate two nodes for the same metabolite within two separate compartments to aid in visualization downstream; however, user data for the given entity would be properly mapped to both nodes.

After the network is curated for the user-specified organism, each node's degree (or magnitude of edges or connections) is determined to aid in the user's downstream ability to avoid visualizing high-degree components, such as a proton or water, on the metabolic network, which can lead to graphical entanglement and cluttering and a decrease in computational performance [cite:Waller;GigaScience;2020].

2. Data overlay and broadcasting for missing entities

In order to overlay user data on the global network, first, user-provided gene expression, protein abundance, and/or metabolite abundances' names are mapped to *Metaboverse* compatible identifiers. For components that *Metaboverse* is unable to map, a list will be returned to the user so they can provide alternative names to aid in mapping. Second, provided data values are mapped to the appropriate nodes in the network. In cases where gene expression data is available, but protein abundance data is missing, *Metaboverse* will take the average (or user-defined??) of the available gene expression values to broadcast to the protein node. For complexes, all available component values (metabolites, proteins, etc.) are averaged (or user-defined??). Nodes for which values were inferred will be marked by a dashed border during visualization to clearly show which values are known and which were inferred.

3. Collapsing reactions with missing expression or abundance user data

After data mapping is complete, *Metaboverse* will generate a collapsed network representation for optional viewing during later visualization. We did, however, choose to enforce a limit of up to three reactions that can be collapsed as data down a pathway should only be inferred so far. We also

enforced certain parameters for reaction collapse as follows: 1. If a reaction has at least one known or inferred value for inputs (reactants) and one known or inferred value for outputs (products), the reaction will be left as is. During the entire reaction collapse step, known catalysts are included when assessing whether a reaction has measured output values (more of a catalyst should lead to more output in most cases) and inhibitors are included when assessing whether the reaction has measured input values (more inhibitor should lead to accumulation of input in most cases). Catalysts and inhibitors are not included when determining reaction neighbors as described below. 2. If a reaction has at least one known input, the input is left as is, and each reaction that shares the same input with the assessed reaction inputs are determined whether they have a measured output. If the neighbor reaction does not contain an known output value, the reaction is left as is. If the neighbor reaction does contain a measured output, the original reaction's inputs and the neighbor reactions outputs are collapsed to form a single, pseudo-reaction between the two. If the reaction has at least one known output, the inverse is performed where neighbors with identical components as the reactions inputs are assessed for whether a collapsed reaction can be created. 3. If a reaction has no measured values, it is determined if the neighboring reactions on both sides (one sharing the reaction's inputs and other sharing the reaction's outputs) have measured values. If both neighbors contain a measured value, a collapsed pseudo-reaction is created summarizing all three reactions.

For pseudo-reactions, appropriate notes are included to describe the collapse. During visualization, these pseudo-reactions are marked by black dashed edges and dashed node borders. A graphical representation of how this reaction collapse is performed can be found in [Figure 2](#).

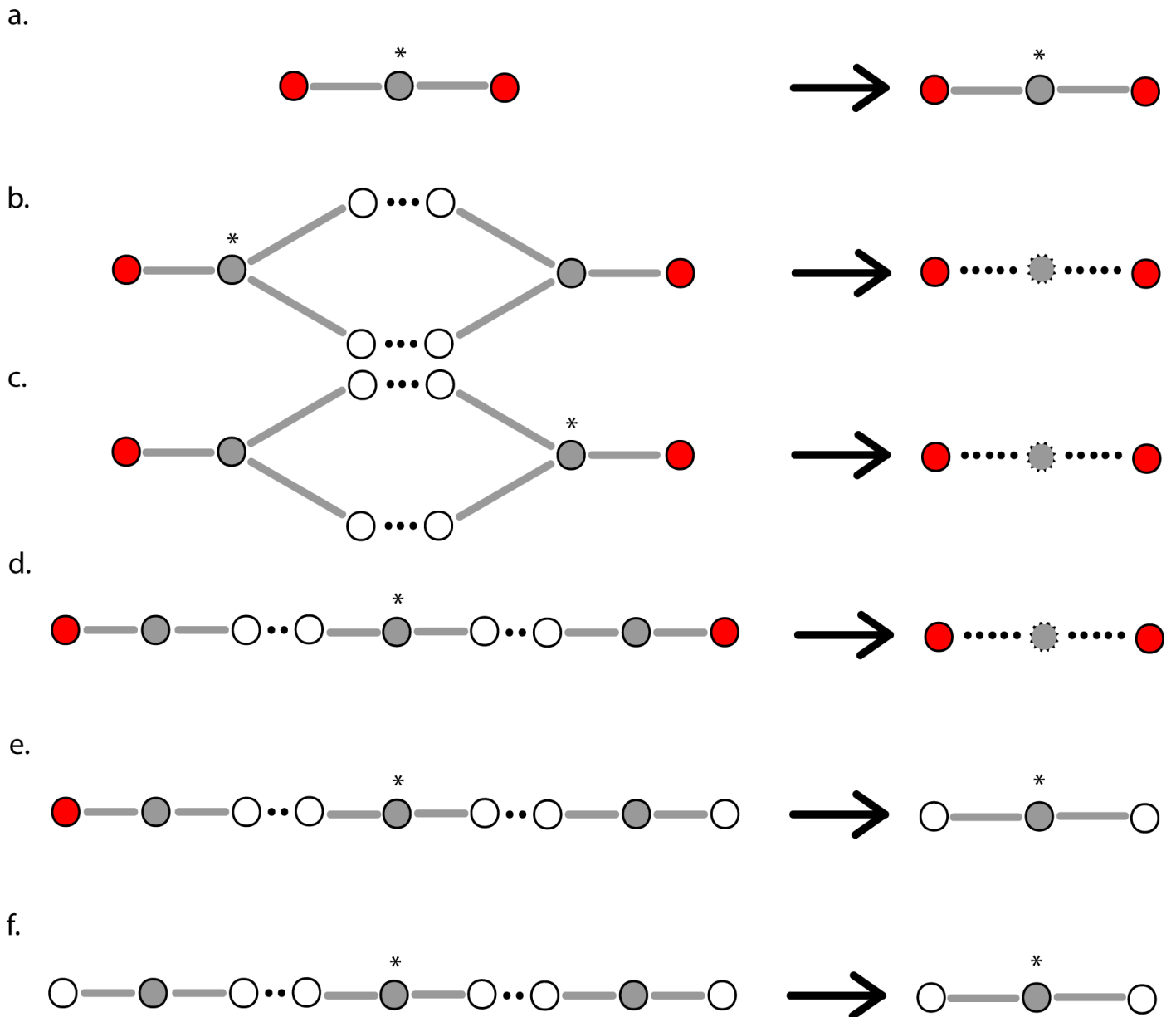


Figure 2: Reaction node collapse schematic. (a) For reactions where at least one input and at least one output component contain a measured value from the user data, the reaction will be maintained as is. (b) Where an input of a reaction is known but no output has a known value, *Metaboverse* will search for all neighboring reactions that contain identical inputs. If the neighboring reaction has a known output value, the two reactions will be merged into one pseudo-reaction. (c) Where an output of a reaction is known but no input has a known value, *Metaboverse* will search for all neighboring reactions that contain identical outputs. If the neighboring reaction has a known input value, the two reactions will be merged into one pseudo-reaction. (d) For reactions with no known values, neighbor pairs that match the inputs and outputs of the considered reaction will be evaluated for whether their respective outputs and inputs both have known values. If values are known for both neighbors, the three reactions will be merged into one pseudo-reaction. (e) As in (d), but if one neighbor does not contain a value, the one does contain a value, no reaction merging will be performed. (f) As in (d), but if neither neighbors contain known values, no reaction merging will be performed. An asterisk (*) indicates the target reaction being considered for a given reaction collapse. A red node indicates a reaction input or output with a measured value. A white node indicates a reaction input or output with no measured value. A grey node indicates a reaction. A grey node with a dashed border indicates a pseudo-reaction. A solid edge indicates a known relationship. A dashed edge indicates a relationship inferred via reaction merging.

4. Regulatory pattern (motif) searches

5. Nearest neighborhood searches

In order to visualize all global connections, a user can select an entity (a gene, protein, or metabolite) and visualize all reactions that the component is involved in. By doing so, the user can visualize other downstream effects a change of one entity might have across the global network, which consequently aids in bridging and identifying any motifs that may occur between canonically annotated pathways. These neighborhoods can be expanded to view multiple downstream reaction steps and their accompanying genes, proteins, and metabolites by modulating the appropriate user option in the app.

Users can also limit which entities are shown by enforcing a degree threshold. By setting this value at 50, for example, the graph would not show nodes that have 50 or more connections. One caveat, however, is that this will occasionally break synchronous pathways into multiple pieces if one of these high-degree nodes was the bridge between two sides of a pathway.

6. Network visualization and exploration

6.1 Dynamic network plotting

Users interact with *Metaboverse* through an interactive app interface. The app uses Electron, a cross-platform app framework that uses JavaScript, HTML, and CSS to design the interface. *Metaboverse* thus comes packaged as a single executable app with all necessary dependencies included for running on Linux, MacOS, and Windows.

Interactive graphing is handled using the D3 and JQuery JavaScript libraries. Force-directed graphs are constructed by taking the user selection for a pathway or entity and determining the reactions that are components of that pathway. All inputs, outputs, modifiers, and other components of these reactions, along with edges where both source and target are found in the sub-graph as nodes, are included and plotted. Relevant metadata, such as user-provided data and reaction descriptions, can be accessed by the user in real time. Metadata for categorical displays, such as edge or node type, are extracted from the metadata during graphing of the sub-network.

Some performance optimization features are included by default to prevent computational overload. For example, nearest neighbor sub-graphs with more than 1500 nodes, or nodes with more than 500 edges will not be plotted as plotting of this information in real-time is computationally prohibitive.

6.2 Visualizing pathways and super-pathways

In order to visualize a pathway, a user selects their pathway of choice and all component reactions and their reactants, products, modifiers, and metadata are parsed from the global network. Super-pathways help categorize these pathways and are defined as any pathway containing more than 200 nodes.

6.3 Visualizing compartments

Compartments are derived from Reactome annotations. Compartment visualizations are generated using D3's hull plotting feature. Compartment boundaries are defined at the reaction levels and made to encompass each reaction's reactants, products, and modifiers for that given compartment.

6.4 Annotations

Annotations for each reaction are derived from the Reactome database. Pseudo-reactions annotations do not include this information; instead they include notes on which reactions were collapsed to create the selected pseudo-reaction. All inferred pseudo-reactions and protein or complex values are displayed with dashed edges to differentiation from measured values.

6.6 Additional features

While Metaboverse will continue to undergo development and new features will be added, we will briefly highlight some additional features available at time of publication. We encourage users to check the documentation for more current updates and information regarding use of *Metaboverse* [docs].

6.6a Toggle genes

As gene components can crowd the graph space, users can toggle gene display on and off using the appropriate button. The graph is then refreshed to either include or ignore gene components based on their node meta-tag.

6.6b Toggling values

Users can switch between coloring nodes based on the value or statistic provided by toggling the appropriate button. Colorbar information for the dataset is saved in the graph metadata during curation and used to generate a colorbar. The colorbar for statistics is represented using a $-\log_{10}$ scale for a statistic value originally ranging between 0 and 1.

6.6c Toggling features/labels

By default, reaction and feature labels are displayed by hovering the mouse over the node. Reaction or feature nodes can have the labels statically displayed by selecting the appropriate button. An event watch function is used to watch for this user selection and update the display of the node labels.

6.6d Toggling collapsed reactions

By selecting the appropriate button, users can toggle between displaying a full or collapsed pathway representation of the sub-network. By selecting this button, the graph is refreshed using the appropriate reaction dictionary, where for graphing of the collapsed representation, a reaction with available pseudo-reactions substituted for the original reactions are included for graph propagation.

6.6e View curated pathway image

While *Metaboverse* graphs networks dynamically, users may be more familiar or comfortable with classical, curated pathway layouts when exploring their data. For a given pathway graph, the user can select the appropriate button and *Metaboverse* will open a new window with the Reactome curated pathway layout.

6.6f Saving graphs

Users can generate a PNG output file for any network created in *Metaboverse* by selecting the appropriate button.

6.6g Nearest neighbor and hub thresholding

The number of nearest neighbors to graph, or the limit to number of edges a graphed node can have, can be modulated by the user using the appropriate input spaces. When graphing a nearest neighbors network, *Metaboverse* will recursively fetch related reactions and their neighbors until a node display threshold is reached. This allows the user to visualize downstream effects of a change that may propagate across several reactions. The hub threshold option prevents plotting of nodes with more than the specified number of edges. This is handling during graphing by excluding any entity nodes that meet this criteria as the neighborhood is propagated. This is particularly useful in removing hub nodes, such as water or protons, which may be less relevant to the user experience and can quickly clutter the graph. This feature can also help plot more extensive neighborhoods, as often neighborhoods quickly link to high-degree nodes, such as water, and limit graphing ability.

6.6h Metadata display

To help inform the user of selection information and relevant metadata, a space in the legend bar during visualization is reserved for spaces where this information can be displayed, which is updated based on the user's input as it is provided.

7. Packaging

The *Metaboverse* app is packaged using Electron. Back-end network curation and data processing is performed using Python and the NetworkX library. Front-end visualizatin is performed using Javascript and relies on the D3 and JQuery packages. Saving network representations to a PNG file is performed using the d3-save-svg and string-pixel-width packages. Documentation for Metaboverse is found at metaboverse.readthedocs.io. Continuous integration services are performed by Travis CI to routinely run test cases for each change made to the *Metaboverse* architecture. The *Metaboverse* source code can be accessed at <https://github.com/Metaboverse/metaboverse>. The code used to draft and revise this manuscript, as well as all associated scripts used to generate and visualize the data presented in this manuscript can be accessed at <https://github.com/Metaboverse/manuscript>.

[dependencies table]

Name	Reference
HTML	
CSS	
Javascript	
Electron	
JQuery	
D3	
string-pixel-width	
d3-save-svg	
Python	
pandas	
numpy	

Figure 3: Author contributions. Table visualizing respective contributions of each author.

References
