

Harnessing the Power of Scientific Python to Investigate Biogeochemistry and Metaproteomes of the Central Pacific Ocean

Noelle A. Held^{§‡}, Jaclyn K. Saunders^{‡§}, Joe Futrelle[‡], Mak A. Saito^{‡*}

<https://youtu.be/WYmAu0GiSU4>

Abstract—Oceanographic expeditions commonly generate millions of data points for various chemical, biological, and physical features, all in different formats. Scientific Python tools are extremely useful for synthesizing this data to make sense of major trends in the changing ocean environment. In this paper, we present our application of scientific Python to investigate metaproteome data from the oxygen-depleted Central Pacific Ocean. The microbial proteins of this region are major drivers of biogeochemical cycles, and represent a living proxy of the ancient anoxic ocean. They also provide a look into the trajectory of the ocean in the face of rising temperatures, which cause deoxygenation. We assessed 103 metaproteome samples collected in the Central Pacific Ocean on the 2016 ProteOMZ cruise. This data represents ~60,000 identified proteins and over 6 million datapoints, in addition to over 6,600 corresponding chemical, physical, and biological metadata points.

An interactive data analysis tool which enables the scientific user to visualize and interrogate patterns in these large metaproteomic datasets in conjunction with hydrographic features was not previously available. Bench scientists who would like to use this oceanographic data to gain insight into marine biogeochemical cycles were at a disadvantage as no tool existed to query these complex datasets in a visually meaningful way. Our goal was to provide a graphical visualization tool to enhance the exploration of these complex dataset; specifically, using interactive tools to enable users the ability to filter and automatically generate plots from slices of large metaproteomic and hydrographic datasets. We developed a Bokeh application [BOKEH] for data exploration which allows the user to hone in on proteins of interest using widgets. The user can then explore relationships between protein abundance and water column depth, hydrographic data, and taxonomic origin. The result is a complete and interactive visualization tool for interrogating a multivariate oceanographic dataset, which helped us to demonstrate a strong relationship between chemical, physical, and biological variables and the microbial proteins expressed. Because it was impossible to display all the proteins at once in the Bokeh application, we additionally describe an application of Holoviews/Datashader [HOLOVIEWS], [DATASHADER] to this data, which further highlights the extreme differences between oxygen rich surface waters and the oxygen poor mesopelagic. This application can be easily adapted to new datasets, and is already proving to be a useful tool for exploring patterns in ocean protein abundance.

Index Terms—oceanography, microbial ecology, biogeochemistry, omics, visualization, bokeh, datashader, holoviews, pandas, dask, jupyter

§ Massachusetts Institute of Technology, Cambridge, MA
 ‡ Woods Hole Oceanographic Institution, Woods Hole, MA
 * Corresponding author: msaito@whoi.edu

Copyright © 2018 Noelle A. Held et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Introduction

Oceanography is concerned with understanding the ocean as a holistic and dynamic system, integrating information from disciplines such as biology, chemistry, geology, and physics. But just how to incorporate this multivariate data is a key challenge in the field. For example, research expeditions commonly generate millions of data points, all with different formats, scales, and primary research goals. Scientific Python tools can help oceanographers synthesize multivariate information to make sense of trends; here we present an application to investigate metaproteome data from the oxygen poor central Pacific Ocean.

The tropical Pacific Ocean contains a naturally low-oxygen region called an oxygen minimum zone (OMZ) (Figure 1). Biological and chemical processes in the OMZ are different from surrounding oxygenated waters. For example, nitrification (use of ammonia or other organic nitrogen sources to fuel processes that typically use oxygen, in simplified form the reaction $\text{NH}_4 \rightarrow \text{NO}_2 \rightarrow \text{NO}_3$) is a key process in the OMZ but not present in oxygenated waters [ULLOA2012]. OMZs may represent a living proxy of the past anoxic ocean. They are also a picture into the future. Climate change driven by anthropogenic carbon dioxide emissions is causing ocean waters to be warmer and more stratified. This leads to deoxygenation processes and predicted expansion of OMZs [WRIGHT2012]. Thus, understanding the biogeochemistry of existing, natural OMZs is important for predicting conditions in the future ocean.

We travelled to the oxygen poor Pacific ocean in winter 2016 to study biological and chemical processes on the ProteOMZ research cruise (<https://schmidtocean.org/cruise/investigating-life-without-oxygen-in-the-tropical-pacific/>). To explore the biogeochemical processes in this region, we collected over 103 metaproteomics samples at various locations and depths, representing 56,577 identified proteins and over 6 million individual data points. In addition, we collected over 6,600 corresponding chemical and physical metadata points (18 variables) which provide context to the biological protein data. Proteins are the molecular machines driving biogeochemical transformations within microbial cells; as such, protein datasets provide a rich look at ecosystem function. To our knowledge this is the largest marine metaproteomics dataset to date.

In this paper, we describe our efforts to create an integrated proteomics and metadata visualization tool. The application is

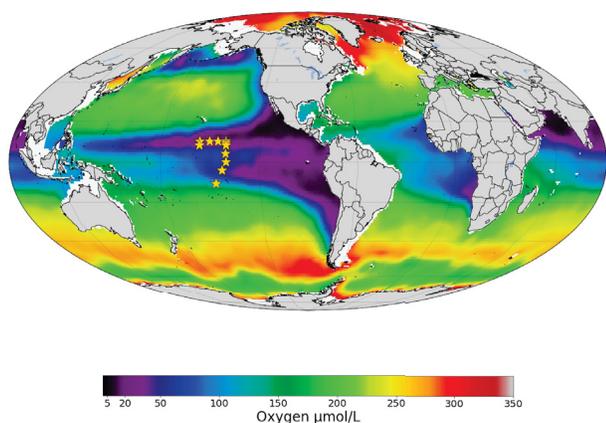


Fig. 1: Oxygen concentrations in the world ocean at 300m depth. Warm colors indicate more oxygen, cool colors indicate less. The sampling locations of the ProteOMZ cruise are overlaid as yellow stars. ProteOMZ samples the oxygen-depleted tropical Pacific region. Oxygen data: World Ocean Atlas [GARCIA2014]

intended as an exploratory tool for user-driven discovery of patterns in oceanographic protein abundance in relationship to hydrographic and ecological context. Using Bokeh [BOKEH] as the main visualization library, we developed an application that integrates multivariate data into interactive plots and tables. We begin by describing the data model, which emerged both from the inherent properties of the data and the constraints of the Bokeh library. We then describe an example in which we demonstrate major phylogenetic and functional differences between oxygen rich surface waters to oxygen poor mesopelagic waters. Due to performance constraints, the Bokeh application can only display a subset of the data. Therefore we additionally describe an application of Datashader implemented in Holoviews and Jupyter Notebook to visualize patterns in the entire dataset. This notebook further demonstrates functional partitioning between oxygen rich and poor waters, emphasizing the extremity of these biogeochemical differences. We conclude with a brief discussion of the benefits and drawbacks of our data construction and library choices, as well as some recommendations for developers and scientists working with these libraries.

Methods and Results

In situ sampling and data acquisition

Samples were collected in January-February 2016 at 14 locations (stations) in the tropical Pacific ocean. At each station, large volume in situ pumps were deployed at multiple depths in the water column. For each pump, hundreds of liters of water were passed through stacked 51 μM , 3 μM and 0.2 μM filters. The data described here is for the 0.2-3 μM filter range which includes most single cell phytoplankton and free living heterotrophic bacteria. More detail on proteomics analyses can be found in [SAITO2014]. The full sample collection and analysis methods for this dataset in particular will be reported in an upcoming publication.

Visualizing Hydrographic Data

We developed a visualization platform to explore the hydrographic data, which includes physical parameters such as temperature

ProteOMZ EXPEDITION 2016

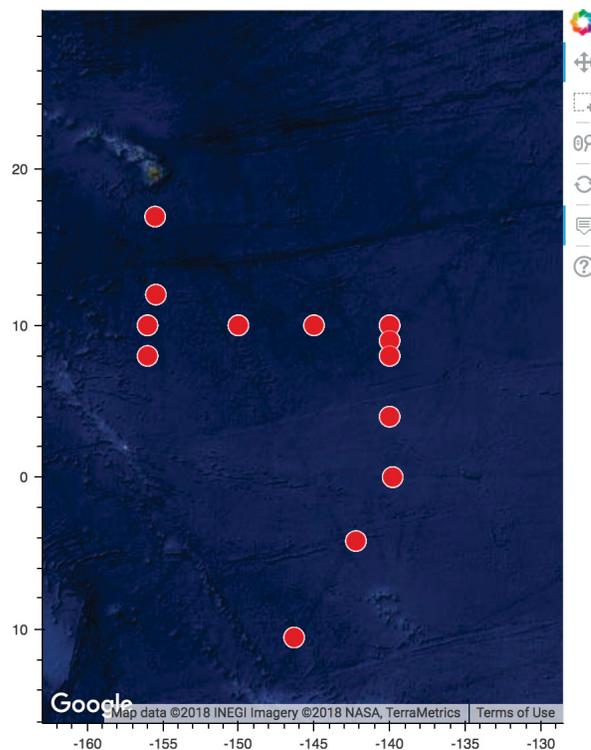


Fig. 2: Station map for the hydrographic data.

and chemical parameters such as ammonium concentrations. The visualization was written with Bokeh in the the Jupyter Notebook interface and produces a standalone html document as the output. This allows the document to be shared with colleagues and, importantly, does not require them to have bokeh or even python installed on their machine. The visualization consists of a map rendered in Google Maps using the `gmap` function in Bokeh (Figure 2a) and scatter plots showing the vertical distribution of the hydrographic parameters throughout the water column, with surface values at the top (Figure 2b). The plots are arranged with `gridplot`. This visualization is fed from a hydrographic data CSV file, where the data for each variable is in a separate column. This facilitates ingestion into Bokeh's `ColumnDataSource`, allowing the plots to be linked. Thus, when the user selects data from one plot, corresponding data for that location is highlighted in the other plots.

Bokeh Application

The main product of this work is a fully interactive Bokeh server application, which integrates protein quantitative data, protein annotations, and hydrographic data. For full interactivity among plots, Bokeh requires data to be in a single 2D `ColumnDataSource`. Thus, the first challenge we faced was how to compress our multidimensional data into a 2D format that could be accessed by multiple plots and updated via widgets. The protein quantitative data is a CSV formatted output which is generated directly from the common proteomics analysis program Scaffold [SCAFFOLD]. For illustrative purposes in this paper we use a truncated CSV file containing 15,000 of the nearly 60,000 identified proteins. However, we have had success using the entire 60,000 protein dataset.

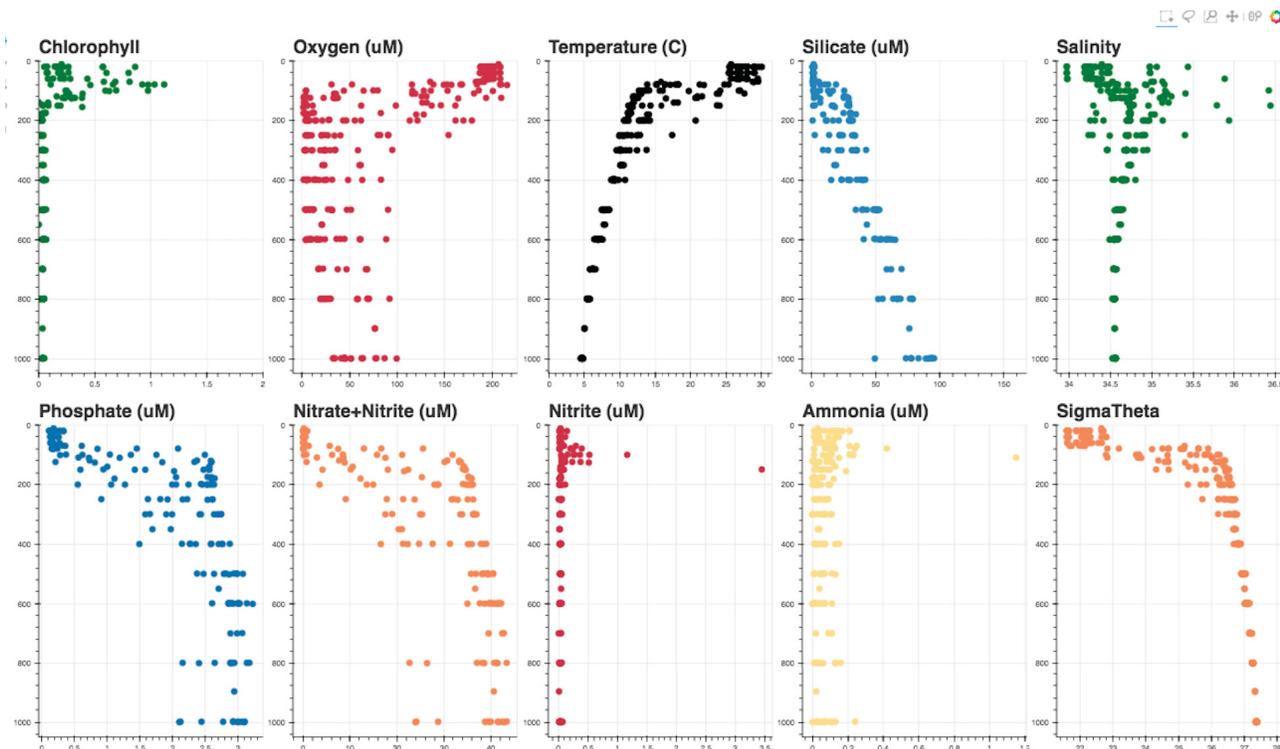


Fig. 3: Hydrographic data as a function of water column depth. The file is exported by Bokeh as a standalone html document, allowing it to be easily shared with collaborators.

The CSV file is read as a pandas dataframe [PANDAS] and consists of 103 rows (one per each unique sampling location and depth) and over 15,000 columns, where each column represents a different protein that was identified in the field sample. This 15,000 protein dataset is a subset of the full protein dataset of 60,000 proteins. The protein annotation information is read as a separate file and includes taxonomic and functional information about each protein in the dataset. Finally, the hydrographic data consists of 103 rows, again, one per each unique sampling location and depth) and 16 columns each containing a hydrographic or chemical parameter also measured on the expedition. We combined all three of these dataframes into a combined data model, allowing the entire application to be fed from ColumnDataSources generated from slices of a single Pandas dataframe (Figure 4). This facilitates connectivity among the plots via tools such as hover and tap, and allows the user to explore all the visualizations using widgets for protein annotation and hydrographic data.

We now describe a use case to demonstrate the utility of the application (Figure 5-8). On initial load, the user can see a map of the ProteOMZ 2016 sampling locations (Figure 5). The user can select a Station via a widget and display a vertical distribution of all of the proteins identified at this station throughout the water column, from surface to deep. Hovering over a protein in the vertical distribution profile displays its identity. The vertical distribution, protein annotation table, and protein vs. hydrographic data charts are directly linked since they are fed through the same ColumnDataSource. Selecting a protein via the TapTool highlights it in the vertical profile, protein annotation table, and in the Protein vs. Hydrographic data chart. A user who is interested in a specific protein can select it from the table, which updates the vertical line profile to highlight that protein. For instance, we can select the most abundant protein in the dataset at Station 5

and see that it is a nitrate oxidoreductase protein (Figure 6). The protein vs. hydrographic data chart displays protein abundance as a function of various hydrographic features, which can be selected by a widget. With the hydrographic widget we select nitrate (NO_3), a product of nitrification, and see that abundance of nitrate oxidoreductase is positively correlated with nitrate (Figure 7). The protein is negatively correlated with its reactant ammonium (NH_4), and also with the intermediary product nitrite (NO_2). Consistent with the idea that nitrification is prevalent in oxygen minimum zones, we see that the protein is negatively correlated with oxygen (O_2) concentrations.

Selecting a station additionally populates a vertical profile of the total number of unique proteins identified (line) and number of peptide-to-spectrum matches expressed on a log scale (bubble) at each depth sampled. In proteomics, we do not measure proteins but instead parts of proteins called peptides, which are then matched to spectra that are predicted in silico from a genome database. The peptide-to-spectrum match indicates the total number of peptides identified (non unique). Typically the number of peptide-to-spectrum matches is related to the number of unique peptides identified; we see this reflected in the data at Station 5. For instance, we see that at depths 200m and below there are more proteins and more peptide-to-spectrum matches than in surface waters. However, though the number of unique proteins is approximately constant between 200 and 500m, the number of PSMs varies.

So far we have looked only at protein function, but a user may also be interested in taxonomic origin of the proteins. At Station 5, we see in the Diversity of Microbial Proteins bar graph that most of the proteins we identified are from the group “Other Bacteria,” which encompasses most heterotrophic bacteria including the nitrifying bacteria (Figure 8). There are also many

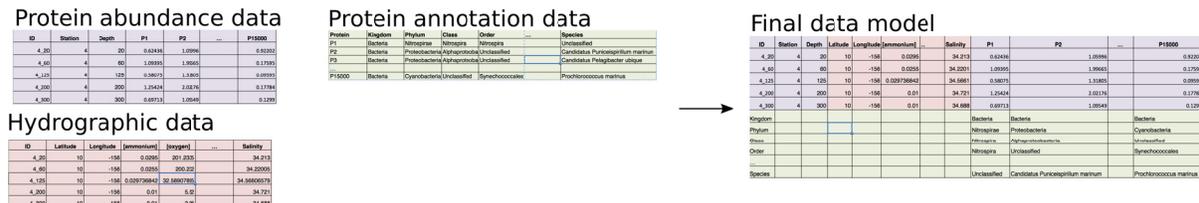


Fig. 4: Data model for integrating protein abundance, protein annotation, and hydrographic data into a single Bokeh ColumnDataSource, allowing for interactivity among the visualizations in the application.

Prochlorococcus and Pelagibacter proteins in the dataset, which is consistent with the fact that these cells are among the most abundant in the ocean [EGGLESTON2016]. A user can select a specific taxon with the taxon widget; for example, we can select “Prochlorococcus” from the taxon widget and redisplay the data (Figure 5). We can now see that Prochlorococcus, a photosynthetic cyanobacterium, is present primarily in the sunlit surface waters above 120m. If we display “Other Bacteria,” we can see that indeed that the heterotrophic nitrifying bacteria are highly abundant in the oxygen-depleted waters beginning around 200m. Thus with just a few clicks we can explore major taxonomic and functional regimes throughout the oxic and suboxic water column.

Application of Datashader

We quickly discovered that attempting to display over 15,000 lines on a single Bokeh plot was infeasible. We thus display only the top 5% most abundant proteins but allow the user to adjust this percentage via the Percentile slider. When the application is run via Bokeh server on a single laptop, only the top 5-10% of proteins can be displayed without significantly slowing down the visualizations. This alone is powerful - over 1000 proteins are displayed on the initial load, and the widgets allow the user to hone in on taxa and processes of interest such that meaningful information is still easy to find. However, it is clear that the data is oversampled and that proteins that are especially low abundance such as cell signalling and regulatory proteins are systematically “lost” in this visualization.

We used Datashader implemented in Holoviews and a Jupyter notebook to view the dataset in its entirety to see if major patterns in protein abundance emerge when all 15,000 test dataset lines are displayed. To improve performance in Datashader line, we re-formatted the dataframe to be two columns (x and y values) with each protein/depth set separated by NaNs. The dataframe was converted to a Dask dataframe [DASK] for performance reasons. Though this data model requires us to copy the “Depth” data 15,000 times, the performance improvement in the Datashader aggregation steps make this step worthwhile.

One question we can ask of the data is whether patterns emerge among proteins that are more or less abundant than average. We normalized the protein quantitation data by dividing each column by its average, such that the resulting data represents the fold-change in the protein in relationship to its mean over the entire water column. In the visualization, a value of 1 on the x axis suggests that protein abundance is equal to the mean; below 1 the protein is less abundant than average and above 1 the protein is more abundant.

In the datashader plot, the data is overlaid on itself such that areas with more saturated color indicates a high number of proteins with similar fold-change in concentration (Figure 8). This shows the partitioning of microbial proteins on depth. Proteins

that are abundant in the surface converge to 0, or “disapper” around 120m. At Station 5, the warm sunlight euphotic mixed layer ends at approximately 120m. These surface proteins are most likely attributed to Prochlorococcus, an abundant bacterium that lives only in sunlight waters. Below 120m, proteins attributed to heterotrophic bacteria become abundant.

Discussion

We designed a data integration and discovery tool for the ProteOMZ research expedition. In just a few clicks, the application allows users to explore trends in protein abundance, probe relationships between protein abundance and hydrographic data, and dial in to biological processes of interest. As an example we describe how we were able to rapidly investigate the taxonomic and functional differences between oxygen replete surface waters and the oxygen minimum mesopelagic. Since the application uses data from a common proteomics data file format, it will be simple to plug new oceanographic datasets into this application as they become available.

A key challenge to this project was building a data model that worked most efficiently with the libraries we selected. For instance, the Bokeh ColumnDataSource imposed a 2D structure on our multi-dimensional data. In Datashader we faced a similar issue, in which we discovered that aggregating 15,000 individual lines is prohibitively slow; by simply reformatting the data so the aggregation treats the data as individual points we could significantly improve performance. Learning about the constraints of these libraries was an important step in the process of creating this application, especially because we pushed the limits of the libraries. This required deep reading of user guides, API documentation, and Q/A repositories. We thus have two suggestions - 1) that scientists (and others) understand and carefully consider the data models and preferences of the libraries they plan to use before they begin the project and 2) that documentation of the data models and best practices in data formatting be more explicitly referenced in library user guides and be made easier to understand for the non-expert.

Another challenge we faced were problems with API stability. In large part this is due to the fact that we chose to work with libraries that are still in V0 release. We quickly learned to version control our code and used virtual environments to retain specific package versions. Luckily, since the projects are open source it is relatively easy to find information about recent changes, though this is not without frustration. For instance, the Bokeh application originally contained a donut chart, which has since been deprecated. We look forward to more stable releases of the Bokeh, Holoviews, and Datashader libraries, especially because we are now incorporating some of these visualizations into the upcoming Ocean Protein Portal (<http://proteinportal.whoi.edu/>), a

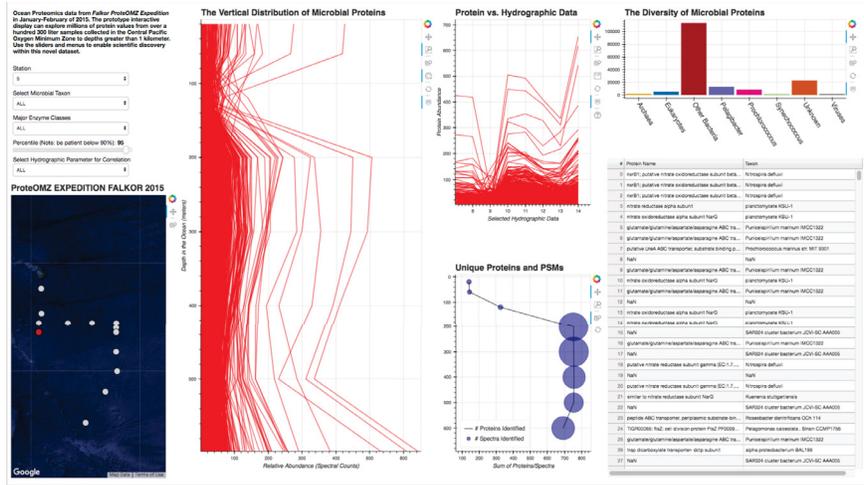


Fig. 5: Initial load of the Bokeh application.

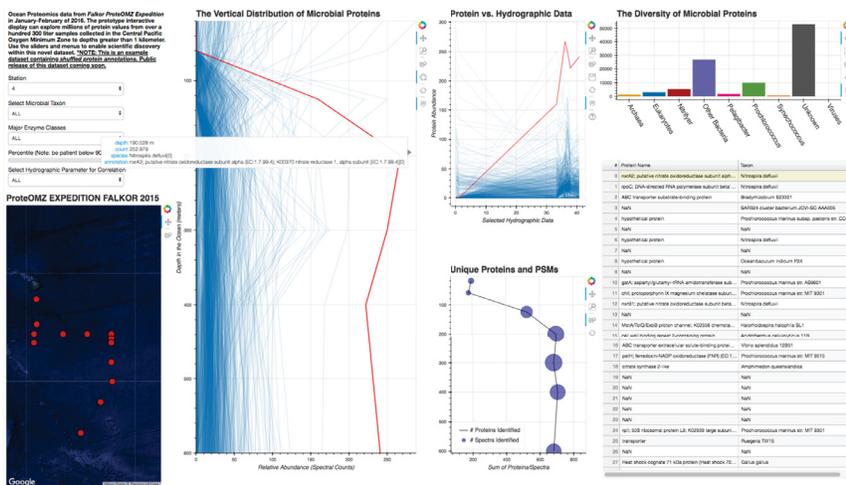


Fig. 6: Selecting on a single protein and investigating relationship to hydrographic data.

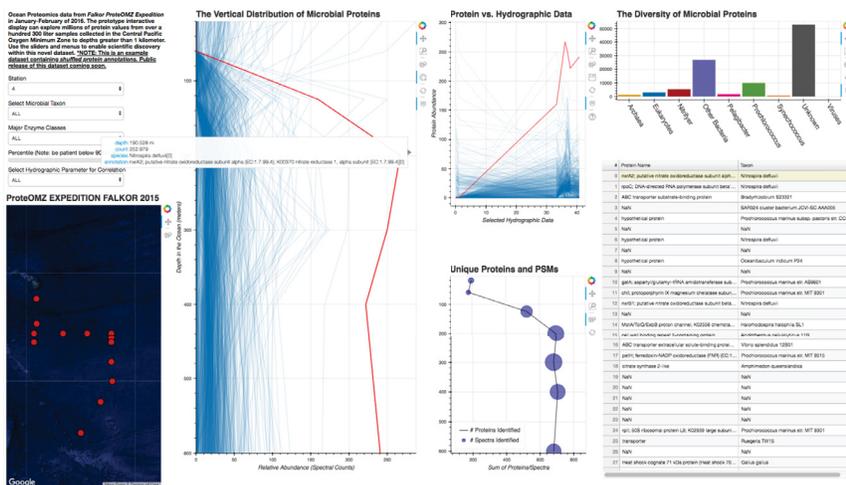


Fig. 7: Filtering on the taxon, we can see that Prochlorococcus proteins are present only in the upper 120m of the water column at this station.

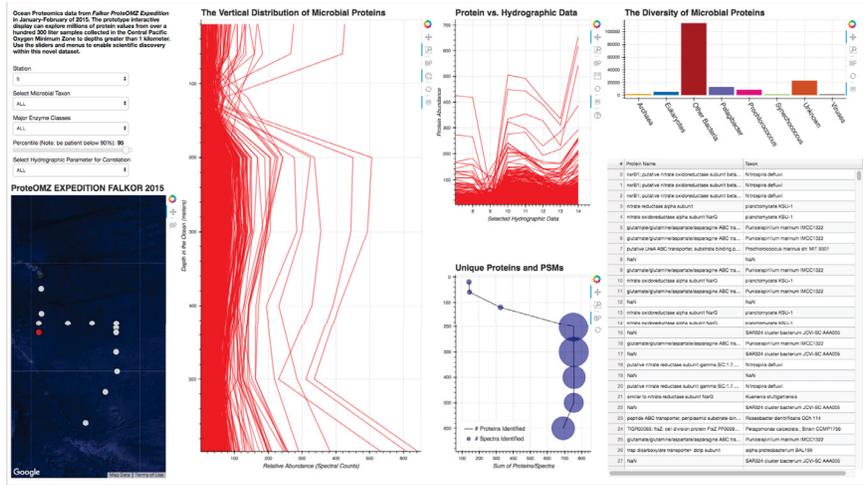


Fig. 8: Selecting “Other Bacteria,” we can see that the nitrifying bacteria become prevalent around 200m in the oxygen minimum zone.

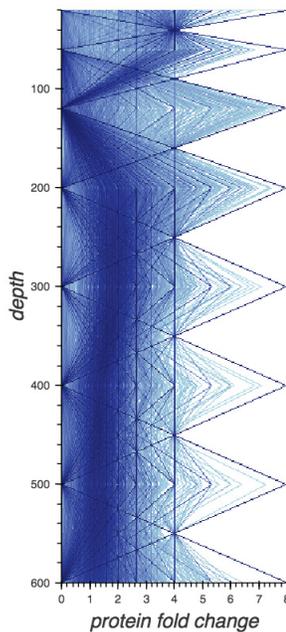


Fig. 9: Datashaded version of the vertical protein distribution plot, displaying all 15,000 proteins at Station 5. Each protein abundance is displayed as the difference from its average, so a value of >1 indicates a protein that is more abundant. A large number of Prochlorococcus proteins is present in the upper 120m; this collection of proteins disappears at the base of the euphotic zone. A large number of proteins is present in approximately the same fold change abundance throughout the mesopelagic region.

data sharing and discovery interface for marine metaproteomics data.

The main benefit of building these visualizations using Scientific Python tools is that scientists who are not primarily programmers can easily manipulate and maintain the code. The code is relatively straightforward, largely due to the fact that the Bokeh and in particular Holoviews backends do much of the heavy lifting. This makes it easier for colleagues to adapt the code to their own datasets. The linked charts in the Bokeh application allow for intuitive (read: more efficient) exploration of the data. In addition, charts generated by Bokeh, Datashader and Holoviews

are beautiful “out of the box.” This is an advantage when we share these visualizations not only with other scientific experts, but also with the general public during outreach events.

The visualizations we built are already proving to be useful. We discuss above just one high level example in which the application helps us to explore taxonomic and functional differences between oxic and suboxic water masses. Finer level analyses are sure to uncover even more exciting trends. We are already plugging in new datasets to the application. As mentioned above, many of these visualizations (in addition to some new ones, such as Holoviews Sankey plot) are being incorporated into the upcoming Ocean Protein Portal, which will make them even more accessible to the scientific community.

Code

Hydrography Visualization: https://github.com/maksaito/proteOM2_hydrography_visualization Bokeh Application: https://github.com/maksaito/proteOM2_visualization_app_public Datashader notebook: https://github.com/naheld/15000lines_datashader

Acknowledgements

This work is supported by a National Science Foundation Graduate Research Fellowship under grant number 1122274 (N. Held) and a NASA Postdoctoral Program Fellowship (J. Saunders). It is also supported by the Gordon and Betty Moore Foundation grant number 3782 (M. Saito) and National Science Foundation grant EarthCube 1639714.

REFERENCES

[BOKEH] Bokeh Project. <http://bokeh.pydata.org/>.
 [DASK] Dask Project. <https://dask.pydata.org/en/latest/>.
 [DATASHADER] Datashader Project. <http://datashader.org/index.html>.
 [EGGLESTON2016] Eggleston, E. M., & Hewson, I. (2016). Abundance of two Pelagibacter ubiqui bacteriophage genotypes along a latitudinal transect in the north and south Atlantic Oceans. *Frontiers in Microbiology*, 7(SEP), 1–9. <https://doi.org/10.3389/fmicb.2016.01534>
 [GARCIA2014] Garcia, H. E., R. A. Locarnini, T. P. Boyer, J. I. Antonov, O.K. Baranova, M.M. Zweng, J.R. Reagan, D.R. Johnson, 2014. World Ocean Atlas 2013, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation. S. Levitus, Ed., A. Mishonov Technical Ed.; NOAA Atlas NESDIS 75, 27 pp.

- [HOLOVIEWS] Holoviews Project. <http://holoviews.org/>.
- [PANDAS] Pandas Project. <https://pandas.pydata.org/>.
- [SAITO2014] Saito, M. A., McIlvin, M. R., Moran, D. M., Goepfert, T. J., DiTullio, G. R., Post, A. F., & Lamborg, C. H. (2014). Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science* (New York, N.Y.), 345(6201), 1173–7. <https://doi.org/10.1126/science.1256450>
- [SCAFFOLD] Scaffold, Proteome Software <http://www.proteomesoftware.com/products/scaffold/>
- [ULLOA2012] Ulloa, O., Canfield, D. E., DeLong, E. F., Letelier, R. M., & Stewart, F. J. (2012). Microbial oceanography of anoxic oxygen minimum zones. *Proceedings of the National Academy of Sciences*, 109(40), 15996–16003. <https://doi.org/10.1073/pnas.1205009109>
- [WRIGHT2012] Wright, J. J., Konwar, K. M., & Hallam, S. J. (2012). Microbial ecology of expanding oxygen minimum zones. *Nature Reviews Microbiology*, 10(6), 381–394. <https://doi.org/10.1038/nrmicro2778>