Bio+MedVis Challenge IEEE VIS
Taming protein beasts through visualization

Laura Garrison* Hauke Bartsch† Stefan Bruckner‡

University of Bergen, Norway
Mohn Medical Imaging and Visualization Centre, Bergen, Norway

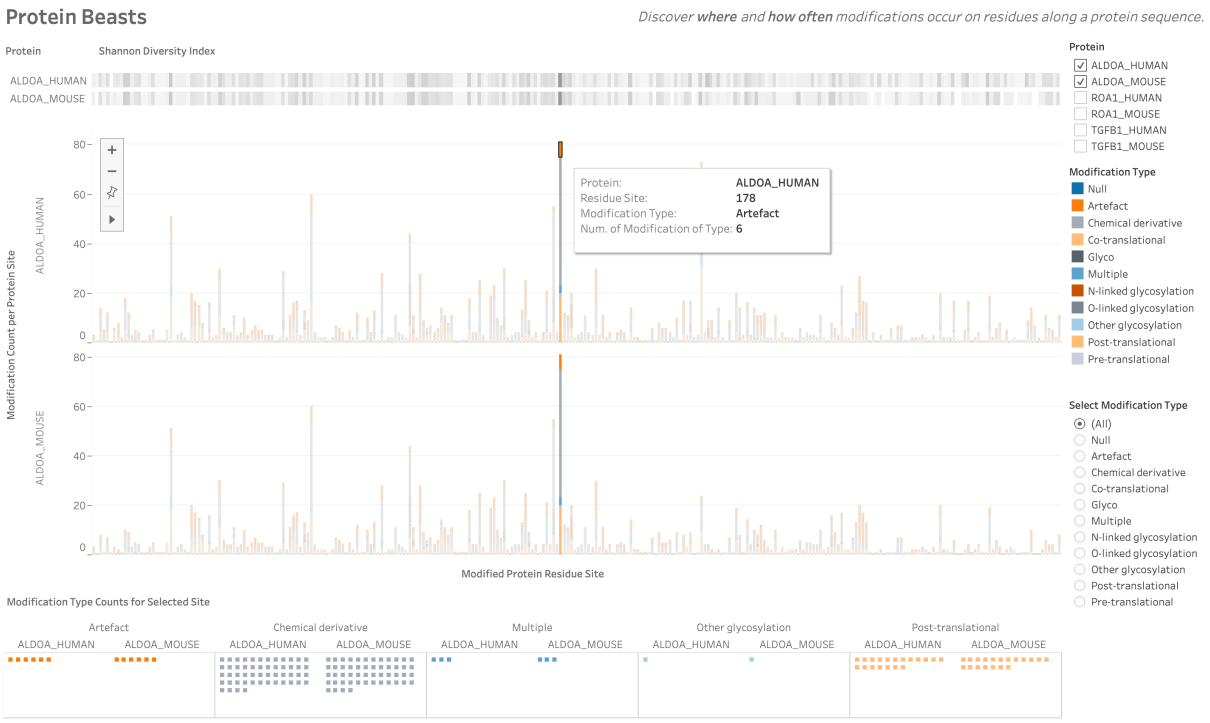

Figure 1: Focus+context interface to facilitate identification and comparison of residues with high occurrences of chemical modifications for human and mouse models. Highlighting interactions in the context views dynamically introduce a focus panel (bottom view) which facilitates comparison of different modifications for each site in a protein sequence within or between species. Subset interactions focus user exploration to a particular modification class (bottom right navigation menu).

Abstract

A key goal in the analysis of proteomic data is to identify and compare the frequency and types of chemical modifications occurring on different residues in a protein sequence. Depicting these data in an uncluttered but informative way is a significant challenge in visualizing these data. For this year's Bio+MedVis Protein Beasts Challenge, we propose a redesign of the existing visualization that reduces visual clutter and facilitates identification and comparison of modifications per site in a given protein. Our design consists of three linked views that facilitate the identification of sites with high modification diversity and which break down the frequency of each type of modification for each residue site in a protein sequence. Highlight and view subset interactions manage visual complexity and focus user attention.

*e-mail: laura.garrison@uib.no
†e-mail: hauke.bartsch@uib.no
‡e-mail: stefan.bruckner@uib.no

Index Terms: Human-centered computing—Visualization—Visualization techniques—; Human-centered computing—Visualization—Visualization design and evaluation methods

1 Motivation

Proteins are essential biological structures in living organisms that are comprised of residues, i.e., linked amino acids. During protein synthesis, chemical modifications, such as post-translational modifications, on these residues can impact the protein's structure and function. Such modifications can have positive, negative, or neutral effects on the protein that more broadly can impact the overall fitness of the organism. Understanding the chemical basis of these modifications, as well as where and the frequency in which they occur, has the potential to help researchers identify and treat rare diseases. Recent efforts to model the occurrence of different chemical modifications across protein sequences have yielded unexpected abundances of modifications at individual residues in a protein sequence that require considered approaches to visualize these data without clutter. For this year's Bio+MedVis Challenge at IEEE VIS, we propose a redesign of the existing visualization, shown in Fig. 2, describing the occurrence and abundance of different chemical modifications in a protein sequence. The main

goals of the visualization are to show experts where modifications are most likely to occur, and to show the occurrence of multiple modifications at a single site. However, this information is difficult to extract with the existing visualization, which is cluttered and uses non-optimal visual encodings for the available screen space and intended user tasks. Our redesigned visualization, shown in Fig. 1, consists of three linked views to facilitate identification and comparison of mutated protein site characteristics at different levels of granularity. An interactive prototype is available at `https://tinyurl.com/mu9656bm`, which we created using Tableau[1].

2 Proposed Redesign

Our key requirements for the redesign include: **[R1]** Declutter the existing visualization, **[R2]** Enable identification and comparison of multiple modifications at a single site, **[R3]** Improve accessibility for colorblind users, and **[R4]** Enable comparison between mouse and human models of a protein.

Our redesign facilitates identification and comparison of the diversity and abundance of modifications at each site in a given protein at three levels of granularity. Fig. 1 provides an overview of the complete interface, which uses a focus+context approach with interactions to enable comparison of modifications per protein site within and between species [R4]. It is designed for use on desktop devices.

To identify the modification sites, we retain the sequential structure of residues from the existing visualization, echoing common practice for visualizing omics sequence data [2] that is familiar to domain experts. Positioning protein residues in a linear, ordered sequence enables users to quickly identify site(s) on which different modifications have occurred [R2]. We exclude unmodified residue sites, similar to the practice of visually filtering out introns in genomic sequence visualizations [6] to remove uninteresting regions from analysis and economize screen space [R2].

Modification frequency is described on three levels. At the coarsest level, shown in the topmost panel of Fig. 1 and in detail in Fig. 6, we summarize per-site modification diversity, i.e., the number of modification types and their proportion, in a given protein sequence. To do this, we use Shannon's diversity index (H), calculated as $H = -\sum p_i \ln(p_i)$, where p_i denotes the proportion of each modification type at a given site, where the index iterates over each type of modification present at that site. A high H means that the site has relatively more different types of modifications present, and that each type of modification occurs in similar amounts [R2]. Conversely, a low H indicates a low diversity site with likely only one type of modification present, and in low amounts. H is mapped to luminance with a greyscale colormap from white ($H = 0$) to black (max. H in a protein sequence). Luminance has been shown perceptually effective in understanding ordered magnitudes [5], and is more space efficient than a position-based encoding. This mapping produces a linear sequence of greyscale tiles organized according to the protein's sequence, echoing the appearance of PCR gels. The visualization may be sorted according to H in ascending or descending order, as in Fig. 7.

The next level showing modification position and frequency of each modification type for each site is shown in the middle panel, and uses a stacked bar metaphor. Bar charts are considered highly effective in characterizing distribution and identifying extrema [5, 8]. Our bar metaphor is also inspired by work by Schatz et al. [9], who use bars to show the time that a ligand was bound to certain residues of a protein. Stacking the bars captures the overall modification frequency on a site, as well as the proportion of each type within a site [4] [R1-R3]. Subset options permit exploration of the per-site frequency of a single modification type, as shown in Fig. 9. This solves perceptual issues that can arise when estimating relative frequencies of modifications that are not on the same baseline. Each modification type is mapped to a hue in an unordered, colorblind-accessible colormap [R3]. Hue is an effective channel to encode categorical attributes, relative to, e.g., shape [5]. This view may also be sorted according to mutation frequency, as in Fig. 8.

Our most fine-grained view, shown in the bottom panel of Fig. 1, focuses on the modification types and their frequencies for a single site which has been selected from either of the two coarser-grained context views. We use a unit visualization metaphor to emphasize the individual, countable occurrences of modifications within each site. This encoding is inspired by prior work on unit and icon visualizations such as by Park et al. [7] and Haroz et al. [3]. We use a simple square glyph to encode each occurrence of a modification to reduce visual complexity [1], and we retain the colormap from the stacked bar view.

[1] https://www.tableau.com/

3 Conclusion

We propose a redesigned visualization of residues modified through posttranslational chemical changes in a protein sequence that reduces the visual clutter through more space-efficient visual encodings than the original visualization. Our design consists of three views that provide information at different levels of granularity to support identification and comparison of per-site modification diversity and frequency between human and mouse protein models. Although our current design excludes the three-dimensional structure of proteins, this could be incorporated with a linked structural view to explore connections between sites, e.g., those with similar diversity values, that appear unconnected in our two-dimensional view.

Acknowledgments

This research is supported by the University of Bergen and the Trond Mohn Foundation in Bergen (#813558, Visualizing Data Science for Large Scale Hypothesis Management in Imaging Biomarker Discovery (VIDI)). Parts of this work have been carried out in the context of the Mohn Medical Imaging and Visualization Centre (MMIV) and the Center for Data Science (CEDAS) in Bergen, Norway.

References

[1] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. *Comput Graph Forum*, pp. 39–63, 2013.

[2] L. A. Garrison, I. Kolesar, I. Viola, H. Hauser, and S. Bruckner. Trends & opportunities in visualization for physiology: A multiscale overview. *Comput Graph Forum*, 41(3):609–643, 2022. doi: 10.1111/cgf.14575

[3] S. Haroz, R. Kosara, and S. L. Franconeri. Isotype visualization: Working memory, performance, and engagement with pictographs. In *Proc ACM Human Factors in Computing Systems*, pp. 1191–1200, 2015. doi: 10.1145/2702123.2702275

[4] L. Howorko, J. M. Boedianto, B. Daniel, et al. The efficacy of stacked bar charts in supporting single-attribute and overall-attribute comparisons. *Visual Informatics*, 2(3):155–165, 2018. doi: 10.1016/j.visinf.2018.09.002

[5] T. Munzner. *Visualization analysis and design*. CRC press, 2014.

[6] S. Nusrat, T. Harbig, and N. Gehlenborg. Tasks, techniques, and tools for genomic data visualization. In *Comput Graph Forum*, vol. 38, pp. 781–805, 2019. doi: 10.1111/cgf.13727

[7] D. Park, S. Drucker, R. Fernandez, and N. Elmqvist. Atom: A grammar for unit visualizations. *IEEE Trans Vis Comput Graph*, 24(12):3032 – 3043, 2017. doi: 10.1109/TVCG.2017.2785807

[8] B. Saket, A. Endert, and Ç. Demiralp. Task-based effectiveness of basic visualizations. *IEEE Trans Vis Comput Graph*, 25(7):2505–2512, 2018. doi: 10.1109/TVCG.2018.2829750

[9] K. Schatz, J. J. Franco-Moreno, M. Schäfer, A. S. Rose, V. Ferrario, J. Pleiss, P.-P. Vázquez, T. Ertl, and M. Krone. Visual analysis of large-scale protein-ligand interaction data. In *Comput Graph Forum*, vol. 40, pp. 394–408, 2021. doi: 10.1111/cgf.14386

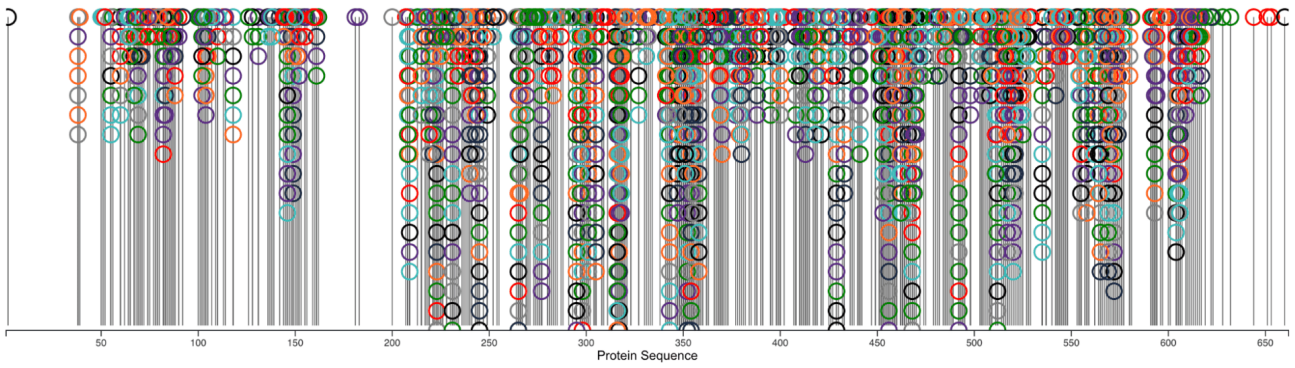

Figure 2: Original visualization for re-design

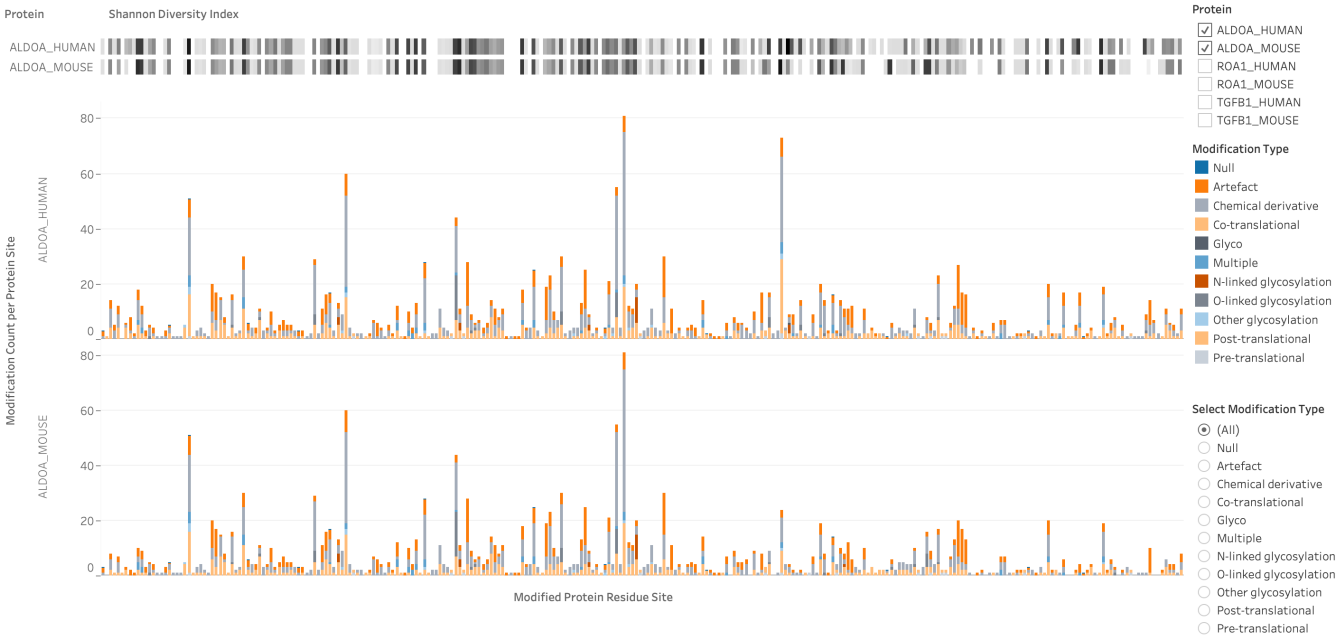

Figure 3: Detail of context panel with Aldoa protein selected for human and mouse models, with Shannon diversity index luminance chart and stacked bar chart describing the counts of each chemical modification per site. Charts share a common x-axis (protein residue site).

Figure 4: Detail of focus panel for the occurrence of modifications at residue site 178 in Aldoa human and mouse proteins, respectively.

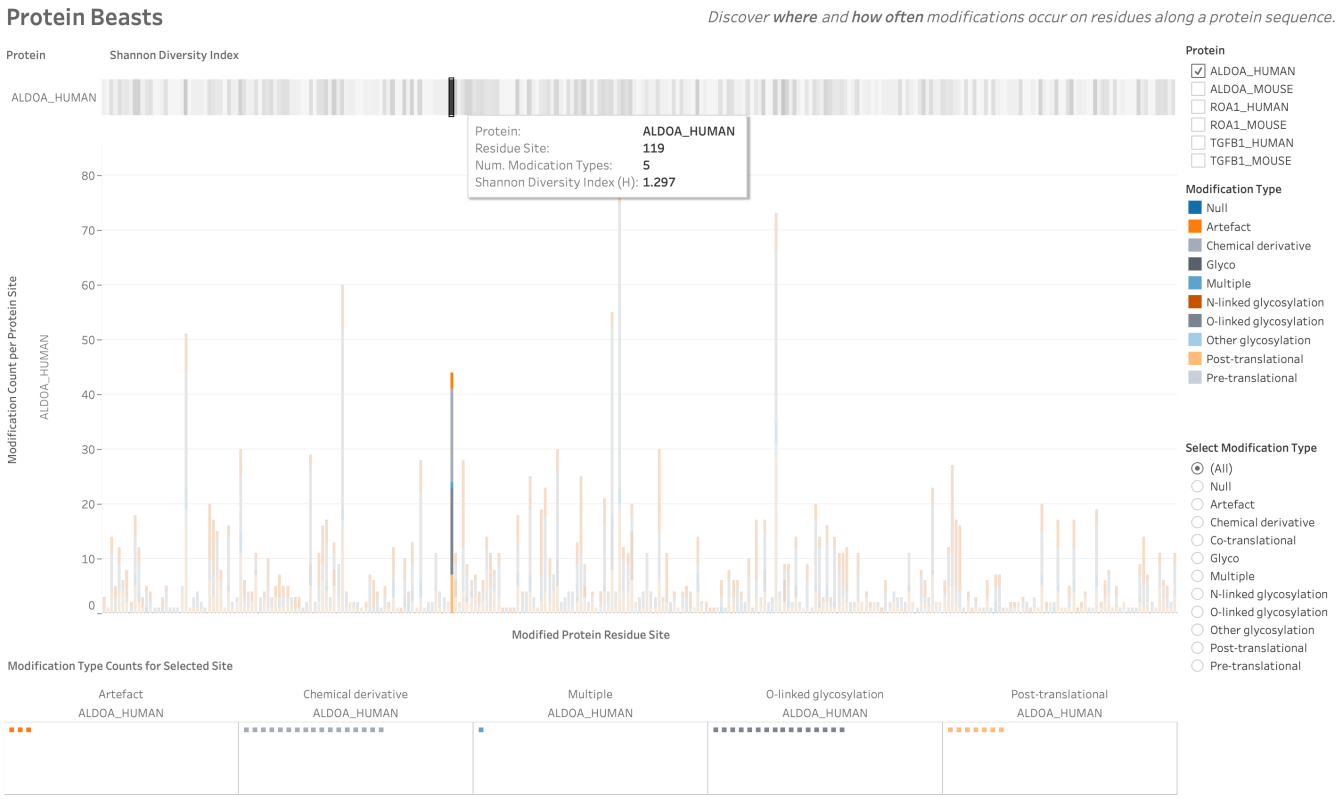

Figure 5: Highlight interaction with tooltip available on hover providing information on the Shannon diversity index at site 119 for the Aldoa human protein. Hovering over this data element highlights the summary of modification types and their frequencies in the stacked bar chart below, and the focus chart at the bottom panel uses a unit visualization to facilitate comparison of each type of chemical modification between different species.

Figure 6: Detail view of view for Shannon diversity index. Orange boxed region indicates a segment where indices differ: the human protein shows a higher richness/evenness of modifications occurring at this site compared to the mouse.

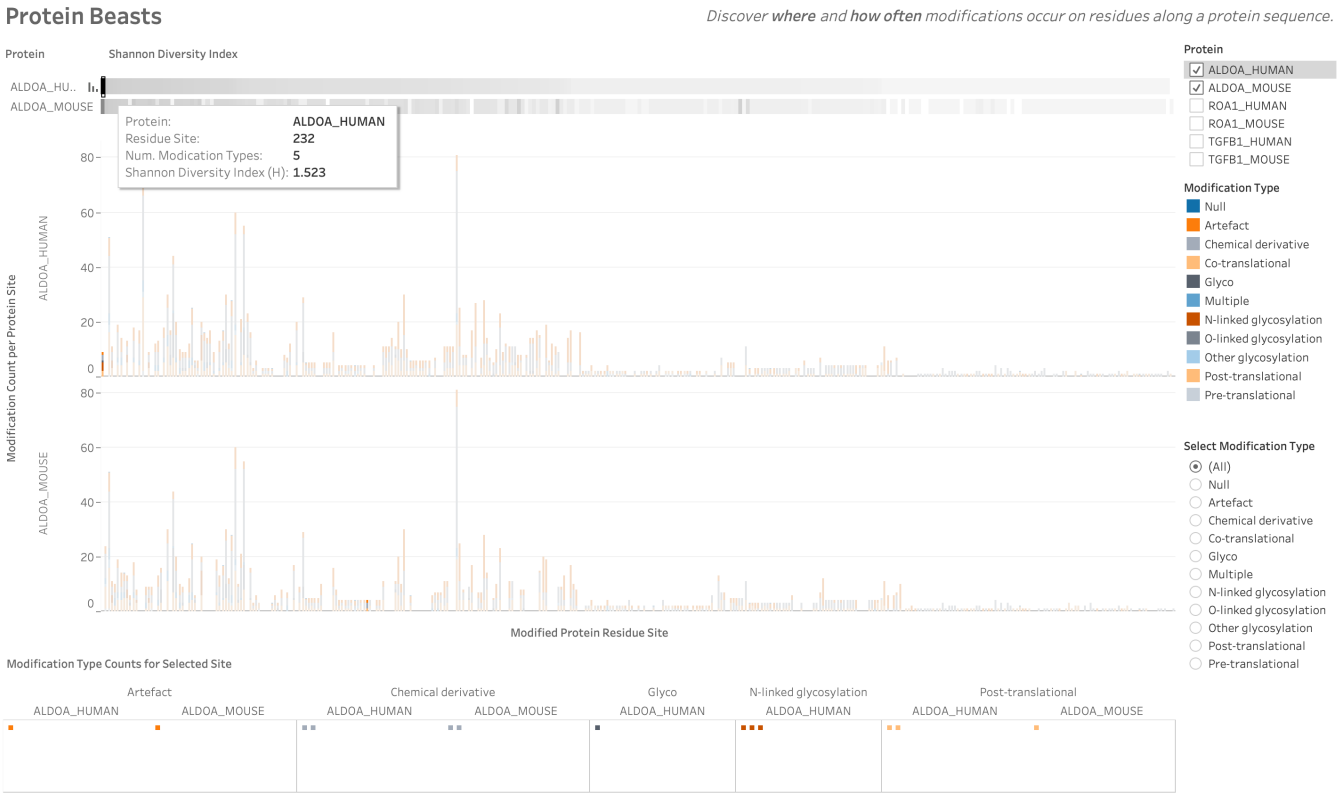

Figure 7: In visualizing the modifications for the Aldoa protein, we can sort the top chart according to Shannon diversity index for the Aldoa human protein, where hovering over a high value highlights the corresponding position in the stacked bar chart and unit visualizations below. Here we have highlighted the highest value, which helps us see that sites with more modification types which are also evenly distributed are more diverse.

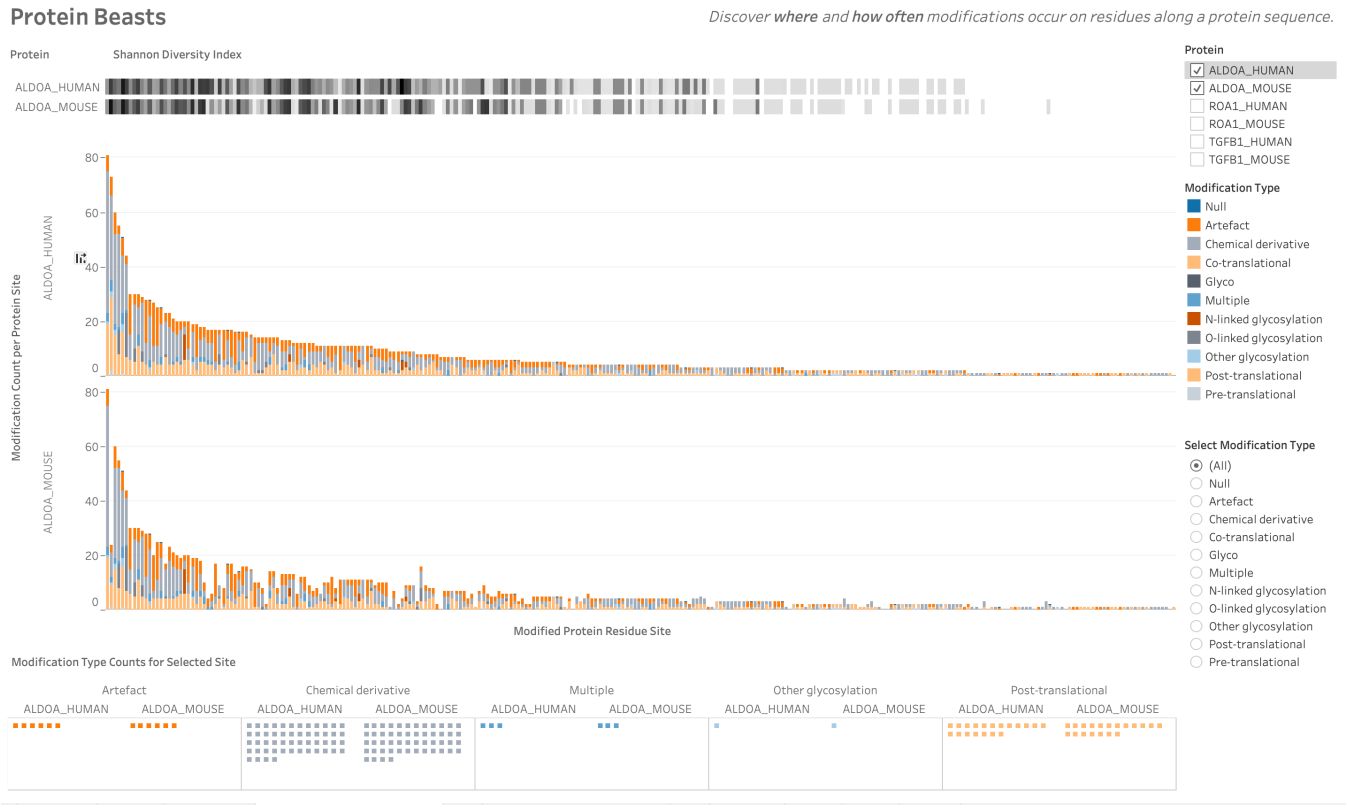

Figure 8: We can also sort the context views by modification frequency (middle stacked bar chart) to understand the role that modification richness and evenness play in establishing the overall diversity of modifications at a site. This chart is sorted according to the overall modification frequency for the Aldoa human protein.

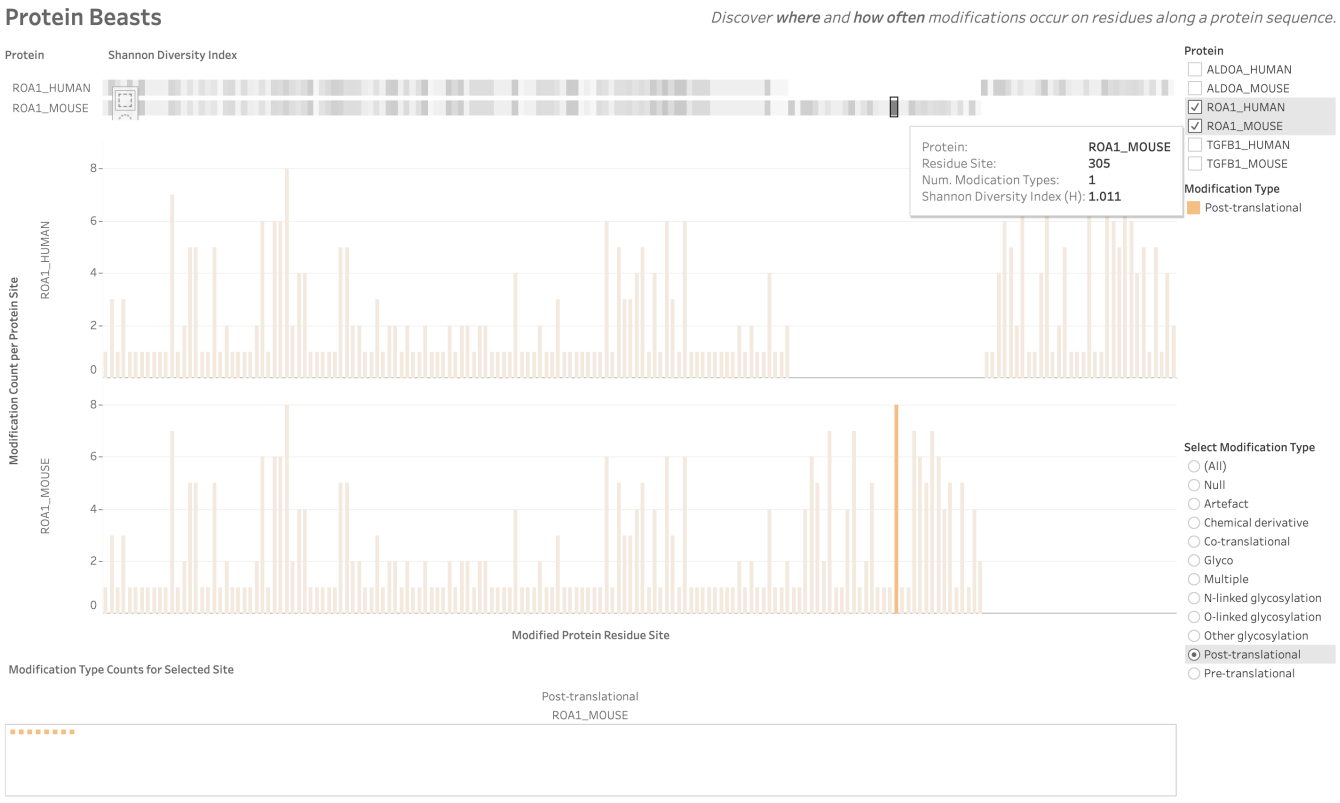

Figure 9: Interactive subset of modification classifications, showing only post-translational modifications for the ROA1 human and mouse proteins. We have highlighted site 305 in this screenshot.