*Bioinformatics*, 38(3), 2022, 849–852
doi: 10.1093/bioinformatics/btab674
Advance Access Publication Date: 29 September 2021
Applications Note

Sequence analysis

# AlphaMap: an open-source Python package for the visual annotation of proteomics data with sequence-specific knowledge

**Eugenia Voytik[1,†], Isabell Bludau[1,†], Sander Willems[1], Fynn M. Hansen[1], Andreas-David Brunner[1], Maximilian T. Strauss[1] and Matthias Mann [1,2,*]**

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany and [2]Department of Clinical Proteomics, NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Olga Vitek

Received on July 30, 2021; revised on September 2, 2021; editorial decision on September 13, 2021; accepted on September 22, 2021

## Abstract

**Summary:** Integrating experimental information across proteomic datasets with the wealth of publicly available sequence annotations is a crucial part in many proteomic studies that currently lacks an automated analysis platform. Here, we present AlphaMap, a Python package that facilitates the visual exploration of peptide-level proteomics data. Identified peptides and post-translational modifications in proteomic datasets are mapped to their corresponding protein sequence and visualized together with prior knowledge from UniProt and with expected proteolytic cleavage sites. The functionality of AlphaMap can be accessed via an intuitive graphical user interface or—more flexibly—as a Python package that allows its integration into common analysis workflows for data visualization. AlphaMap produces publication-quality illustrations and can easily be customized to address a given research question.

**Availability and implementation:** AlphaMap is implemented in Python and released under an Apache license. The source code and one-click installers are freely available at https://github.com/MannLabs/alphamap.

**Contact:** mmann@biochem.mpg.de

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Bottom-up mass spectrometry (MS) has become the leading technology for identifying and quantifying proteomes (Aebersold and Mann, 2003, 2016; Müller *et al.*, 2020). Since peptides rather than intact proteins are measured, visualizing identified peptides and post-translational modifications (PTMs) together with known protein sequence information is an important aspect of downstream MS data exploration. However, the ability to easily integrate and visualize experimental data together with already known sequence annotations is an unmet need in the proteomics community. Although established visualization platforms provide manual visualization of a single experimental sample or dataset at a time (Omasits *et al.*, 2014), there is a lack of tools that support state-of-the-art data analysis software frameworks and that can visualize experimental sequence coverage across multiple samples or datasets in combination with available sequence annotations mined from UniProt, the standard knowledgebase for protein information (Bateman, 2019). To make this wealth of information easily accessible to proteomics researchers, we developed AlphaMap, a Python package that facilitates the visual exploration of peptide-level proteomics data.

## 2 The AlphaMap computational framework

In line with other recently developed software tools from our lab (Strauss *et al.*, 2021; Willems *et al.*, 2021), we implemented AlphaMap in pure Python because of its clear, easy to understand syntax and the availability of excellent supporting scientific libraries. To read fasta files, we leverage the Pyteomics Python package (Goloborodko *et al.*, 2013; Levitsky *et al.*, 2019). Plotly is a well-established plotting library that we use for generating AlphaMap's sequence visualization (Plotly Technologies Inc., 2015), allowing flexible customization and great user interactivity. To enable easy access to the AlphaMap functionality with a low barrier of entry, a stand-alone graphical user interface (GUI) was implemented using the Panel library (Rudiger *et al.*, 2021). AlphaMap can be launched either as a browser-based GUI after simple local installation or as a standard Python module installed via PyPI (Python Software Foundation, n.d.) or directly from its GitHub repository.

In line with the AlphaPept ecosystem (Strauss *et al.*, 2021), we make the AlphaMap code openly available on GitHub, using its many supporting features for unit and system testing via GitHub actions. For code development, we adopted the concept of 'literate programming' (Knuth, 1984), which combines the algorithmic code with readable documentation and testing. Using the nbdev package, the codebase can directly be inspected in well documented Jupyter Notebooks, from which the code is automatically extracted (Kluyver *et al.*, 2016). We envision that these design principles will encourage the broader community to integrate AlphaMap in their own data analysis and visualization workflows with the possibility to easily adopt the code according to specific needs.

## 3 Overview of the AlphaMap workflow

AlphaMap uses peptide-level proteomics data as input. It currently supports the direct import of data processed by MaxQuant (Cox and Mann, 2008), Spectronaut (Bruderer *et al.*, 2015), DIA-NN (Demichev *et al.*, 2020), FragPipe (Kong *et al.*, 2017) and our

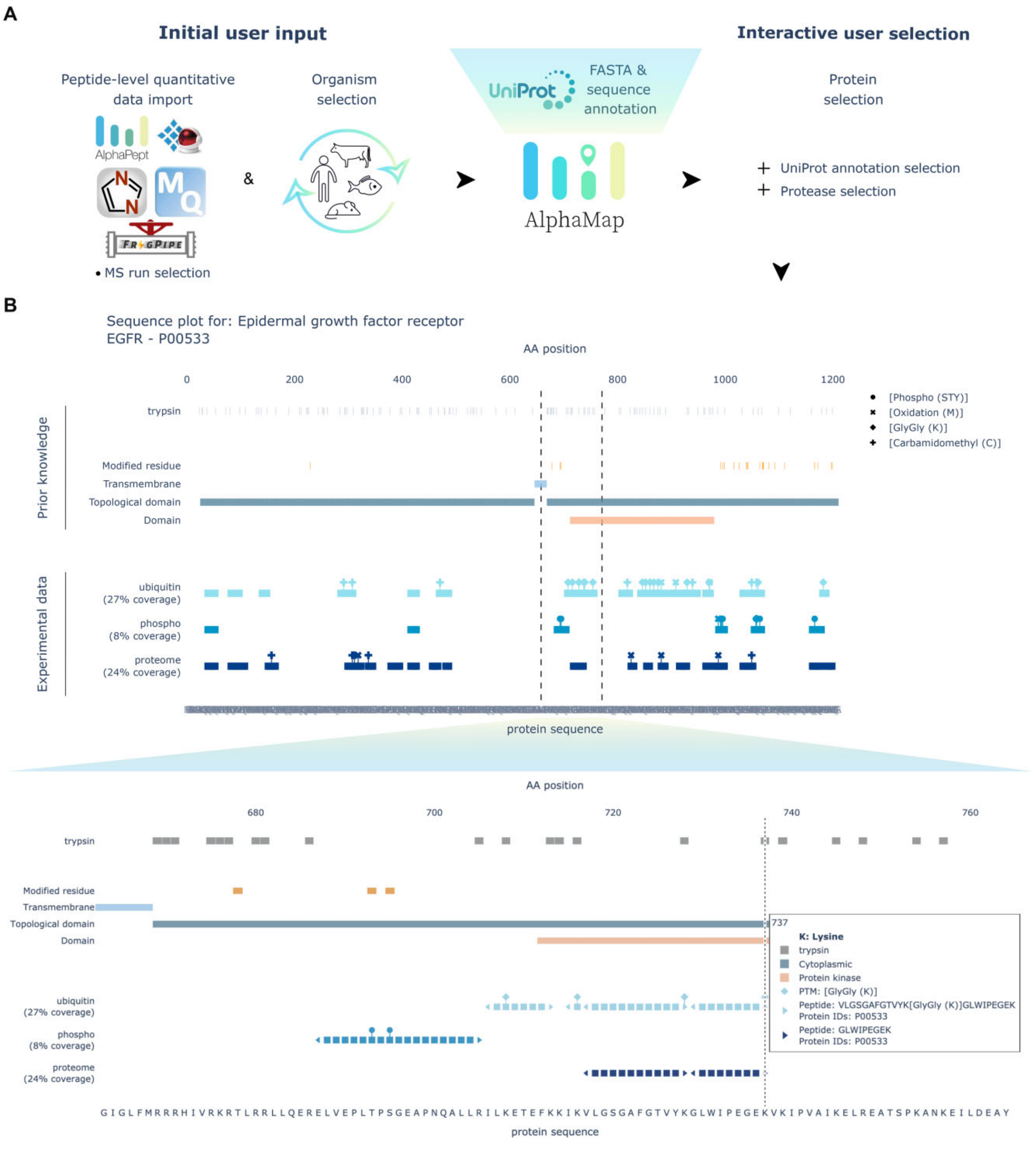

**Fig. 1.** (**A**) Overview of the AlphaMap workflow from MS data upload to the interactive sequence visualization. (**B**) Exemplary sequence visualization for epidermal growth factor receptor (EGFR). A zoom-in on a selected sequence region, indicated by dashed lines, is provided at the lower part of the panel

recently introduced AlphaPept framework (Strauss *et al.*, 2021). In contrast to Protter (Omasits *et al.*, 2014), users can select multiple independent datasets for co-visualization. These could either have been processed by the same or with different MS analysis tools. It is also possible to select only a single sample, or a subset of samples of a given input file for individual sequence visualization. In addition to the peptide-level data generated from LC-MS analysis, AlphaMap leverages a plethora of manually curated sequence-specific protein level information available from UniProt. Fasta files and UniProt sequence annotations are readily accessible in AlphaMap for the 13 most popular UniProt organisms as well as for SARS-CoV and SARS-CoV-2. Functionality to enable the integration of additional organisms is further available as part of our Python package. Finally, the user can select the different layers of information that should be displayed in the interactive sequence representation, including selected protease cleavage sites and UniProt sequence annotations. Figure 1A shows a schematic overview of the AlphaMap workflow. Detailed instructions for its installation and usage are further provided in the supplementary user guide. In addition to interactive sequence visualization of a user-selected protein, AlphaMap provides individual links to external databases and tools for further sequence evaluation in UniProt (Bateman, 2019), PhosphoSitePlus (Hornbeck *et al.*, 2015), Protter (Omasits *et al.*, 2014), PDB (Berman *et al.*, 2000) and Peptide Atlas (Desiere *et al.*, 2006).

## 4 Application of AlphaMap to investigate full proteome and PTM data

Figure 1B shows the sequence visualization of the peptides and PTMs identified for the epidermal growth factor receptor (EGFR) in human A549-ACE2 cells that were infected with SARS-CoV-2 or SARS-CoV (an exemplary viral protein detected in this dataset is visualized in the Supplementary Material) (Stukalov *et al.*, 2021). We show three independent experimental traces: one for full proteome data, one for phospho-enriched peptides and one for ubiquitin-enriched peptides. The proteome data indicates a homogeneous coverage across the entire protein sequence. As expected, phosphorylation and ubiquitination are limited to the C-terminal region of the protein, which is annotated to be exposed to the cytosol. In addition, the kinase domain of EGFR is highly ubiquitinated in our dataset, whereas the surrounding cytosolic regions are phosphorylated. Interestingly, AlphaMap reports that most of our observed phosphorylation sites have been previously identified, whereas none of the identified ubiquitination sites are annotated in UniProt. Please note that unmodified peptides are also observed in both the phospho- and ubiquitin-enriched samples due to the imperfect selectivity of enrichment protocols.

Beyond the uses highlighted here, we envision AlphaMap to facilitate data analysis and interpretation for a variety of different applications:

- Candidate validation: AlphaMap can be used to assess the sequence coverage of identified biomarker candidates (or other proteins of interest) to evaluate possible sequence variations or unexpected anomalies on the basis of readily available sequence information.
- Preparation of panels for publication: Sequence visualizations from AlphaMap can directly highlight the precise MS derived information about proteins of interest in biological or clinical projects.
- Technical comparisons: AlphaMap can be used to evaluate sequence coverage between different data acquisition strategies such as data-dependent and data-independent acquisition, alternative instrument platforms or software tools.
- Optimization of sample processing: Visualization of protein cleavage sites for different proteases can help to optimize sample processing with the goal to achieve a more complete sequence coverage.

## 5 Conclusion

AlphaMap offers an interactive GUI and a Python package for visualizing peptide-level bottom-up proteomics data on the basis of individual protein sequences, including information of curated UniProt sequence annotations and expected proteolytic cleavage sites. We expect that future developments by us and the community will extend the variety of available annotations in AlphaMap, for example by including prior knowledge of sequence conservation or predicted functional domains. In addition, we will integrate quantitative information and differential analysis results into the AlphaMap sequence representations. We envision that AlphaMap will assist MS-based proteomics researchers in inspecting peptide- and PTM-level data, thereby providing valuable information in the process of candidate validation in biological and clinical context.

## Author contributions

I.B. conceptualized the project and together with E.V. and M.M. wrote the manuscript with contributions from all authors. I.B. and E.V. implemented the core AlphaMap functions. E.V. implemented the GUI. S.W. provided important help with the AlphaMap installers. F.M.H. and A.-D.B. provided valuable ideas for the concept and visualization in AlphaMap and F.M.H. further contributed by rigorous testing. M.T.S. designed the general AlphaPept ecosystem and assisted with the nbdev environment. M.M. supervised the study and provided critical feedback on all aspects of the presented work.

## Funding

This study was supported by The Max-Planck Society for Advancement of Science and by the Bavarian State Ministry of Health and Care through the research project DigiMed Bayern (www.digimed-bayern.de, G64b-A1070-2018/131-2 DMB-1805-0008). I.B. acknowledges funding support from her Postdoc.Mobility fellowship granted by the Swiss National Science Foundation [P400PB_191046].

*Conflict of Interest*: none declared.

## Acknowledgements

The authors thank Julia Schessner, Barbara Steigenberger, Jakob Bader and Sophia Mädler for testing and providing critical feedback on AlphaMap. They are grateful to Özge Karayel and Maria C. Tanzer for valuable discussions and for providing experimental data.

## References

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Aebersold,R. and Mann,M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature*, **537**, 347–355.

Bateman,A.; UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*., **47**, D506–D515.

Berman,H.M. *et al.* (2000) The protein data bank. *In Nucleic Acids Res*., **28**, 235–242.

Bruderer,R. *et al.* (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics*, **14**, 1400–1410.

Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol*., **26**, 1367–1372.

Demichev,V. *et al.* (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods*, **17**, 41–44.

Desiere,F. *et al.* (2006) The PeptideAtlas project. *Nucleic Acids Res.*, **34**, D655–D658.

Goloborodko,A.A. *et al.* (2013) Pyteomics – a python framework for exploratory data analysis and rapid software prototyping in proteomics. *J. Am. Soc. Mass Spectrometry*, **24**, 301–304.

Hornbeck,P.V. *et al.* (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.*, **43**, D512–D520.

Kluyver,T. *et al.* (2016) Jupyter Notebooks—a publishing format for reproducible computational workflows. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas – Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016*, Göttingen, Germany, pp. 87–90.

Knuth,D.E. (1984) Literate programming. *Comput. J.*, **27**, 97–111.

Kong,A.T. *et al.* (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods*, **14**, 513–520.

Levitsky,L.I. *et al.* (2019) Pyteomics 4.0: five years of development of a Python proteomics framework. *J. Proteome Res.*, **18**, 709–714.

Müller,J.B. *et al.* (2020) The proteome landscape of the kingdoms of life. *Nature*, **582**, 592–596.

Omasits,U. *et al.* (2014) Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics*, **30**, 884–886.

Plotly Technologies Inc. (2015) *plotly*. Montréal, QC. https://plot.ly (27 September 2021, date last accessed).

Python Software Foundation. (n.d.) Python Package Index – PyPI. https://pypi.org/ (27 September 2021, date last accessed).

Rudiger,P. *et al.* (2021) holoviz/*panel: Version 0.11.3*. doi: 10.5281/ZENODO.4692827.

Strauss,M.T. *et al.* (2021) AlphaPept, a modern and open framework for MS-based proteomics. *BioRxiv*, 2021.07.23.453379. doi:10.1101/2021.07.23.453379.

Stukalov,A. *et al.* (2021) Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. *Nature*, **594**, 246–252.

Willems,S. *et al.* (2021) AlphaTims: indexing trapped ion mobility spectrometry – time of flight data for fast and easy accession and visualization. *Mol. Cell. Proteomics*, 100149.