

An Advanced Database for GlycoPeptide Structure Exploration to build GlycoPeptideStructure Hub: GPS Hub

Michele Casoria^{1,2,3}, Marina Macchiagodena^{1,3}, Carla Bazzicalupi^{1,3}, Claudia Andreini^{1,3,4}, Gianni Cardini^{1,3}, Anna Maria Papini^{2,3}, Marco Pagliai^{1,3}

1 Research Unit of Computational Chemistry, University of Florence, 50019 Sesto Fiorentino, Italy

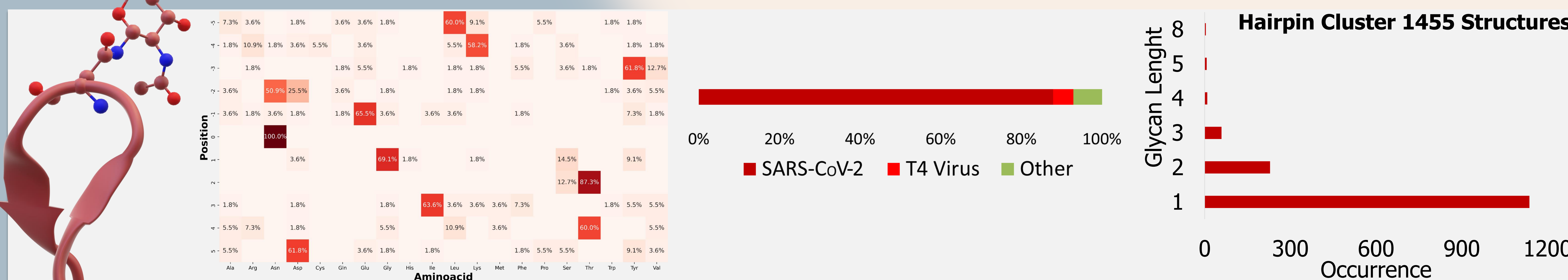
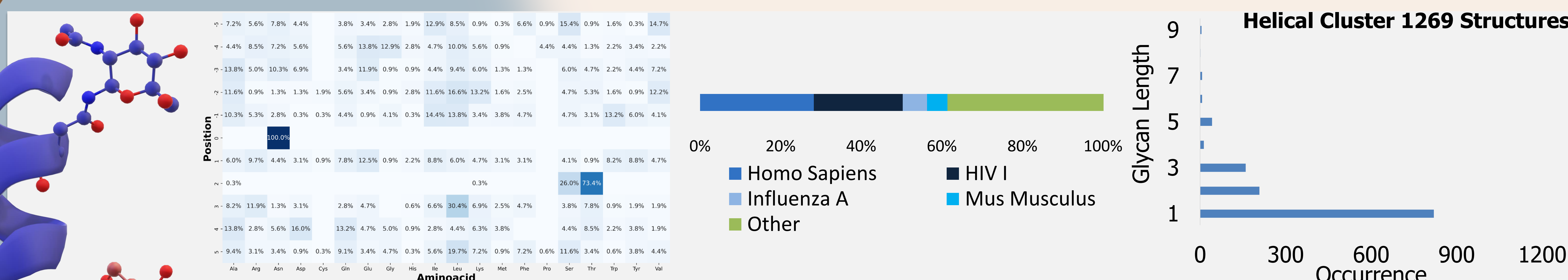
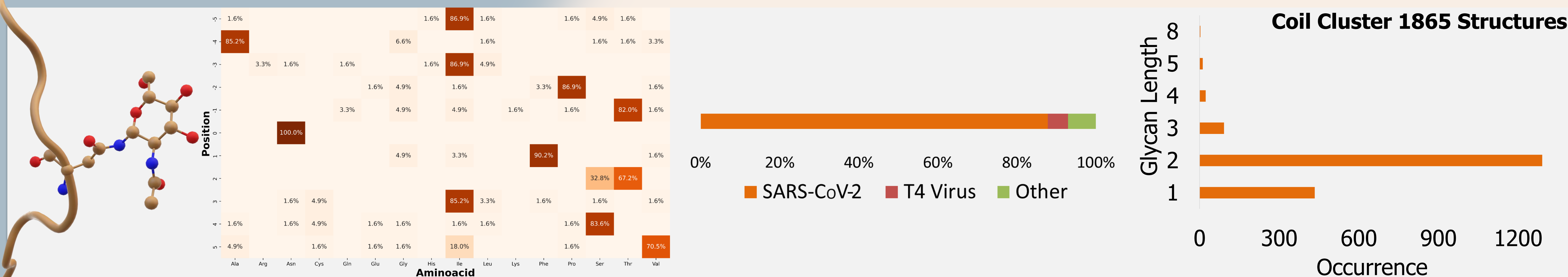
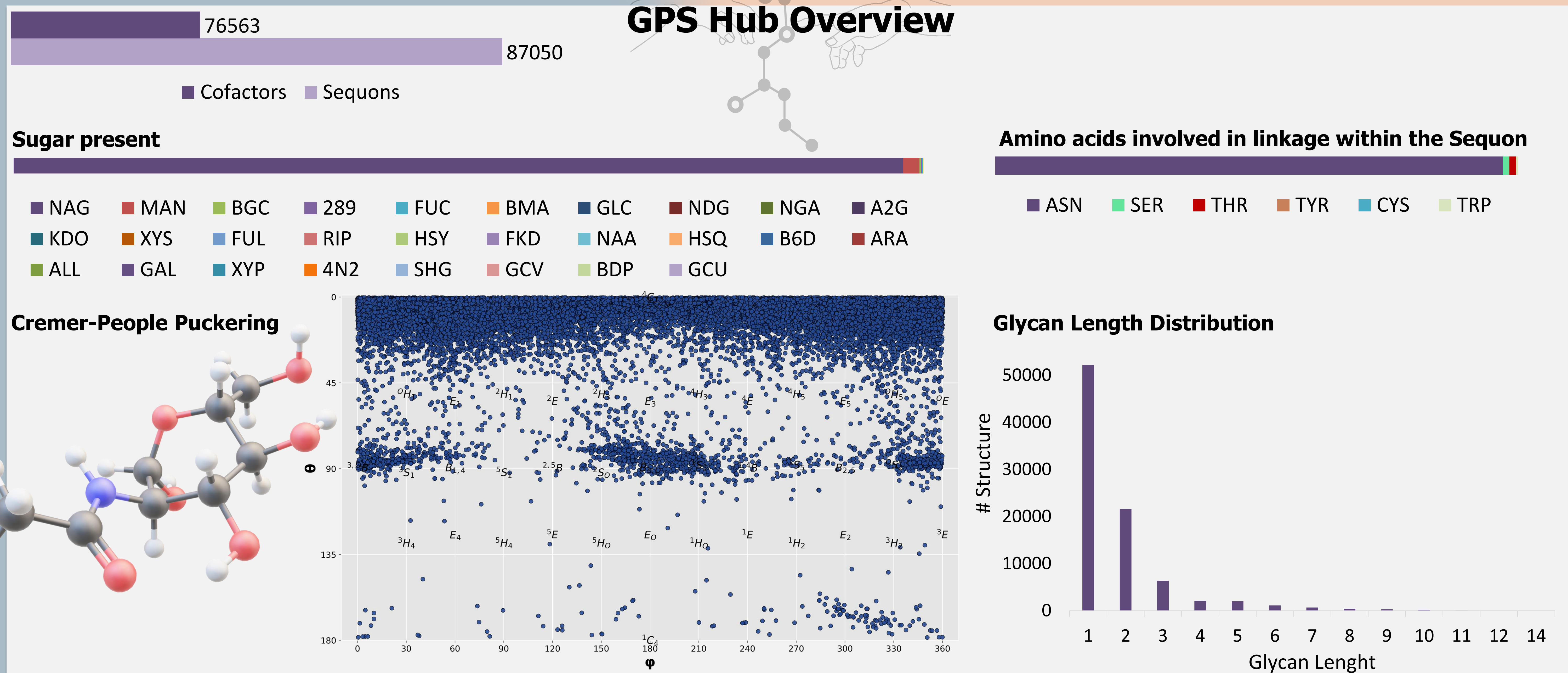
2 Interdepartmental Research Unit of Peptide and Protein Chemistry and Biology, University of Florence, 50019 Sesto Fiorentino, Italy

3 Department of Chemistry "Ugo Schiff", University of Florence, 50019 Sesto Fiorentino, Italy

4 Magnetic Resonance Center - University of Florence, 50019 Sesto Fiorentino, Italy

e-mail: michele.casoria@unifi.it

Using PDB experimental structures, a new database named GPS Hub is proposed. We define a sequon as an 11-mer peptide that spans from positions -5 to +5 relative to the central glycosylation site. We identified 87050 sequons and using machine learning techniques, in particular UMAP and DBSCAN algorithms, we clustered them into 2625 distinct groups. Below, we present the key findings for three of these clusters.



Conclusions

We have developed a database that consolidates comprehensive information on sequon and Glycan structures, starting from experimental data. We employed a machine learning approach to categorize these structures into clusters, uncovering common patterns hidden within glycan sequences.

Acknowledgments