# DiscovEpi: A Tool to automatically retrieve protein data, predict corresponding epitopes and produce potential epitope binding maps for whole proteome.

## Abstract

In the adaptive immune response antigenic peptide presentation by MHC is a major key event to integrate T cells. The binding of a peptide is hereby the most selective step and to understand the principals is of great interest when it comes to drug design. With regards to experimental epitope mapping, it is possible to predict well binding epitopes from an amino acid sequence. Now it would be useful to get an overview about the immunogenicity of every protein of a pathogen’s proteome in silico. From this point any in vitro approach could focus on the putative most immunogenic proteins.

DiscovEpi connects the protein database UniProt and epitope predictor NetMHCpan in a way that one can easily get immunogenicity predictions for whole proteomes. It even creates a visual epitope map of every protein. Individual proteins are being ranked by their density of putative epitopes and the average epitope binding score. An epitope map enables finding epitope-rich regions in a protein and thus paves the way for drug design approaches.

The GUI application is freely available and reduces the user’s effort to get proteome-wide epitope predictions by far. The prediction results are further processed to give the user an estimate of the importance of each protein in the sense of a T cell response. Thereby one can start their in vitro experiments based on these quickly available in silico analysis.

## Introduction

A hallmark of the adaptive immune response is the interaction between cell surface bound MHC class I peptide complexes and the T cell receptor of CD8+ T cells. Binding leads to T cell activation and differentiation into cytotoxic T cells. Moreover, there is evidence that intracellular infection by bacteria such as Staphylococcus aureus is accompanied by presentation of specific epitopes at the cell surface (Bröker et al). Building the complex, selection of the pathogen derived peptide, the epitope, is a key event (Neefjes et al). Additionally, to experimental epitope mapping there are approaches to predict well binding epitopes from a given amino acid sequence for specific MHC molecules like NetMHCpan (Reynisson et al 2020). Given the possibility to predict the binding probability of epitopes for different proteins it is of interest to compare proteins in terms of their immunogenicity and supply the term epitope mapping with an actual visualized epitope map.

In a first step DiscovEpi connects the UniProt protein database with NetMHCpan. One can retrieve protein data for whole proteomes, say all annotated proteins of Staphylococcus aureus and automatically get epitope predictions with rank factors epitope density and average binding score. Finally, DiscovEpi produces a visual of the protein set, where every epitope is highlighted at its respective position.

## Methods

DiscovEpi algorithm is based on Python and the Qt framework using PySide for integration. Protein data from UniProt and prediction data from NetMHCpan are retrieved via RESTful API's from UniProt and IEDB, respectively. Seaborn visualization library is used to produce an epitope map in form of a heatmap. The GUI is created with the Python bindings of Qt and an executable for Microsoft Windows including all packages and dependencies is created with PyInstaller.

The binding score provided by NetMHCpan is converted to a value between 0 and 1 based on the given binding score threshold so that the best binding epitopes have values close to 1 and weak binding ones close to 0.

To rank the proteins there are two values calculated. On the one hand, there is the epitope density describing the ratio of predicted epitopes to the number of possible meres of the respective length. The downside of the density is leaving the binding score of each epitope aside, which gives weight to strong or weak binding epitopes. On the other hand, there is the average score, which cannot differentiate between a few well binding epitopes or many weak binding.

## Results

DiscovEpi is free of charge and enables automatic proteome-wide epitope prediction. It requires least input and manual handling by the user. One has only to keep in mind the UniProt annotation for organism and cellular location to specify the proteome. The IEDB offers different prediction algorithms while to this date only the recommended one is available in DiscovEpi to guarantee state of the art results. Over the last years NetMHCpan has outperformed its competitors and is declared as recommended.

Protein and prediction data are independently processed and put out as XLSX files before passed to produce the visual epitope map created as PNG file.

Comparison of epitope densities at the level of species

The results can be used to compare pathogens according to their immunogenicity overall. Taking the average epitope density of all proteins of a pathogens proteome gives a hint of the immune system's ability to clear out a pathogen. In this case, strains of Staphylococcus aureus have been compared to strains of Influenza A and Coronaviruses. The results indicate that as a mechanism of immune evasion due to long ongoing coexistence Staphylococcus aureus and Influenza A were able to reduce their epitope density compared to Coronaviruses.

Comparison of epitope densities at the level of proteins

Epitope maps allow a closer look at the immunogenicity of proteins. Ordered by their epitope density the heatmap displays every protein of a proteome with highlighted positions where epitopes are located. If there are hundreds of proteins one can cut the list at a given number to get a clear look. Using the visuals, it is possible to identify hotspots of epitopes and thus to identify proteins containing epitope-rich areas. These areas could be of interest when it comes to antigen processing, MHC loading and presentation for T cells. Training of T cells with new identified epitope-rich short sequences of antigens from a pathogen could lead to new targets in e. x. vaccine design.