# **DiscovEpi: Automatic whole proteome MHC-I-epitope prediction and visualization**

**Mahncke C1,2, Schmiedeke F3, Simm S2, Kaderali L2, Bröker B3, Seifert U1, Cammann C1**

1. Friedrich Loeffler-Institute of Medical Microbiology-Virology, University Medicine Greifswald, 17475 Greifswald, Germany
2. Institute of Bioinformatics, University Medicine Greifswald, 17475 Greifswald, Germany
3. Institute of Immunology, University Medicine Greifswald, 17475 Greifswald, Germany

**Summary**

Antigen presentation is a central step in initiating and shaping the adaptive immune response. Pathogen-derived peptides are presented on the cell surface of antigen-presenting cells bound to major histocompatibility complex (MHC) class I molecules. CD8+ T cells that recognize these complexes with their T cell receptor are activated and ideally eliminate infected cells. Prediction of putative peptides binding to MHC class I is crucial to complete the understanding of pathogen recognition in specific immune responses and thus to support drug and vaccine design. Current reliable epitope prediction algorithms focus on the prediction of single epitopes in immunogenic proteins, whereas our developed tool DiscovEpi establish an interface between full proteomes and epitope prediction. The tool allows the automated protein-centric identification of all potential MHC class I-binding peptide epitopes within a proteome adding an overall epitope density and average binding score. DiscovEpi provides a convenient interface between automated multiple sequence extraction by organism and cell compartment from a database like UniProt for subsequent epitope prediction via NetMHCpan. Furthermore, it allows the respective ranking of proteins by their predicted immunogenicity (based on epitope density and average binding score). Visualizing the location of epitopes on the sequence-level can foster the study of adaptive immune responses and support the development of new strategies in vaccine design.

**Availability and implementation**

https://github.com/cmahncke/DiscovEpi

## **Contact**

Clemens Cammann; clemens.cammann@med.uni-greifswald.de

**Introduction**

The adaptive immune response is a complex and tightly regulated defence mechanism that protects the human host from a wide range of pathogens. It is triggered by the interaction of naïve T cells with antigen-presenting cells. In the case of CD8+ T cells, this requires the specific recognition of MHC class I-peptide complexes on the surface of antigen presenting cells by the clonotypic T cell receptors (Neefjes *et al.*, 2011). With the help of cytokines and co-receptors, this leads to activation, differentiation, and clonal expansion of antigen specific cytotoxic CD8+ T cells (CTLs), which subsequently migrate to the site of infection. When they re-encounter the same antigenic MHC class I-peptide complexes on the surface of infected cells, the CTLs induce apoptosis in the target cell through secretion of perforins, granzyme B and/or interaction of Fas with its ligand (Russell and Ley, 2002). Most MHC class I-binding T cell peptide epitopes (MHC-I epitopes with the length of 8-12 amino acids) are generated through degradation of intracellular proteins by the ubiquitin-proteasome system (UPS) in the cytosol (Damgaard, 2021). After their transport into the endoplasmic reticulum (ER) peptides are loaded onto MHC class I molecules and reach the cell surface via the Golgi apparatus (Neefjes *et al.*, 2011; Strehl *et al.*, 2005). To study the involvement of CD8+ T cells in the elimination of a pathogen, it is therefore necessary to analyse the generation of MHC I-epitopes and to determine the exact epitope sequences. Such studies have been successfully performed for viral epitopes, e.g. from Influenza A Virus or SARS‑CoV‑2 (Wu *et al.*, 2019; Da Silva Antunes *et al.*, 2023). In addition, there is evidence that intracellular infection by bacteria such as *Listeria monocytogenes* or *Staphylococcus aureus* is accompanied by presentation of MHC-bound peptides (Pamer *et al.*, 1997; Bröker *et al.*, 2016).

The knowledge of specific epitope sequences can be used for the development of epitope-based vaccines which stimulate immune cells without using the entire pathogen. This targeted approach enhances the efficiency in design and production of the vaccine, reduces possible side effects and increases the coverage if epitopes are shared between different strains of a pathogen (Parvizpour *et al.*, 2020). At stage, development of epitope-based vaccines is based on the manual selection and extraction of potential candidates (Jain *et al.*, 2021) from databases like the UniProt Knowledgebase (UniProt: the Universal Protein Knowledgebase in 2023, 2023), following the accurate independent prediction of single epitopes within the sequence by algorithms like NetMHCpan or MHCflury (Reynisson *et al.*, 2020; O'Donnell *et al.*, 2018). As the protein-centric overall immunogenicity is only indirectly visible identification of immunogenicity and epitope-rich regions within complex proteomes is currently cumbersome and time-consuming.

Our tool DiscovEpi simplified the prediction of protein-centric overall immunogenicity by connecting the UniProt database with NetMHCpan to automize the analysis of whole proteomes or groups of proteins. Especially, the automated extraction of species, strains or even subcellular localization in combination with ranked epitope predictions based on epitope density and average binding score accelerates the research for promising candidates. DiscovEpi visualizes all epitopes at their positions in the respective protein sequences with the intensity based on their binding scores resulting in an user-friendly identification of immunogenic regions and fast comparisons of potential vaccine targets.

**Methods**

DiscovEpi is implemented in Python (van Rossum, 2010) and the Qt framework using the PySide package for integration in python (The Qt Company, Qt for Python project; PySide, version 1.2.4. available at https://pypi.org/project/PySide). For a specified organism and subcellular location contained in the UniProt KB (UniProt: the Universal Protein Knowledgebase in 2023, 2023) amino acid sequences are retrieved and putative epitopes predicted using the neural network NetMHCpan (Reynisson *et al.*, 2020). DiscovEpi normalizes the retrieved binding scores for the set threshold and creates data sheets containing the prediction data and meta information. Finally, DiscovEpi creates a visual epitope map in form of a heatmap. For each amino acid sequence there are grey marks at the positions of its predicted epitopes. The intensity of the mark describes the binding score at the respective position. The binding score at a specific position is calculated as the sum of all binding scores at this position in the amino acid sequence if there are overlapping epitopes.

**Results**

**Workflow and functionality**

DiscovEpi enables the automatic epitope prediction and extends the approach from single protein analysis to broad analysis of protein groups with a shared subcellular localization or even the whole proteome. In this Use Case, DiscovEpi has been used to identify epitope rich regions in membrane bound proteins from SARS-CoV-2. The GUI (graphical user interface) of DiscovEpi (Figure 1A, B, C) allows the automized extraction of all SARS-CoV-2 sequences with specific features. The text fields on the first tab are to specify the set of proteins to retrieve. Beside the name of the organism species or strain, the set can be restricted to a specific subcellular localization as well as to UniProt reviewed proteins only (https://www.uniprot.org/help/organism-name, https://www.uniprot.org/help/subcellular\_location). DiscovEpi output contain an Excel spreadsheet containing meta-data information about the proteins including UniProt ID, protein / gene name, length of the signal sequence (if available) and whole protein sequence, subcellular localization, and the amino acid sequence for each protein (Figure 1A). DiscovEpi can automatically remove redundant sequences before the prediction step. The epitopes will be predicted for every protein sequence retrieved from the UniProt after setting the NetMHCpan parameters (Figure 1B). DiscovEpi forward to NetMHCpan (Dhanda *et al.*, 2019) the HLA allele (default = HLA-A\*02:01), the epitope length (default = 9) and a threshold for the NetMHCpan epitope binding score percentile rank (default = 3.0). Based on the NetMHCpan predictions of single epitopes DiscovEpi calculates the epitope density and average binding score for each protein sequence individually to assess the putative T cell epitope dependent immunogenicity (Supplemental Material 3). Both scores are essential to define research use efficiency as length is a limiting factor for peptide synthesis. The epitope density describes the number of all binding epitopes in relation to all possible peptides of the respective length in the sequence. The average protein binding score focusses only on the retrieved relevant epitopes to gain information about the putative immunogenicity regarding CTLs. For the protein-centric visualization of the potential epitopes DiscovEpi allows the user to setup up the visualization area (Figure 1C) and creates a heatmap of the dataset of protein sequences with their epitopes (Figure 1D). The heatmap is scaled on the maximum protein length (default is the length of the longest sequence in the set) and the number of top ranked protein sequences (default = 50). The proteins to be shown on the map are ordered by the calculated epitope density, so that increasing the vertical resolution through limiting the number of proteins automatically puts a focus on the proteins of highest epitope density.

**Use-Case: Prediction of potential SARS-CoV-2 epitopes**

For the use-case we retrieved 10 unique reviewed protein entries associated with the organism SARS-CoV-2 and the localization term “Membrane” (Figure 1A, Supplemental Material 2). Prediction for epitopes of the length “9” for the allele “HLA-A\*02:01” ended up with a total of 521 epitopes (threshold of NetMHCpan >= 3, Figure 1B, Supplemental Material 3).

The protein length varied between 43 amino acids from the *NS7B\_SARS2* to 7096 amino acids of the Replicase protein *R1AB\_SARS2* (Supplemental Material 2). To ensure good resolution of the smaller proteins we limited the length to 1300 amino acids to still display the full glycoprotein *SPIKE\_SARS2* which is already used in vaccines (Polack *et al.*, 2020) at the expense of truncated sequences of the two replicase proteins *R1A\_SARS2* and *R1AB\_SARS2* (Figure 1D). The resulting epitope map shows wide regions of several bands of mediocre intensity which indicate long sequence parts containing continuous and few overlapping epitopes e. g. for the proteins *VEMP\_SARS2* (position 4-66) and *AP3A\_SARS2* (position 69-120). On the other hand, there are shorter dark grey and black bands indicating that these regions contain relatively many overlapping epitopes or a few epitopes with great epitope scores e. g. for the proteins *VEMP\_SARS2* (position 16-24) and *ORF9C\_SARS2* (position 57-65). In contrast, especially the spike protein *SPIKE\_SARS2* contains regularly distributed epitope regions. To evaluate their immunogenic potential and efficacy in vaccines these regions must be further investigated regarding e. g. hydrophobicity *in vitro*.

As the results clearly indicate the immunogenic potential of a set of proteins and the sites of interest, we could compare these to other pathogens like Influenza A (strain A/Puerto Rico/8/1934 H1N1). Using the same DiscovEpi settings apart from organism, we saw in their 5 membrane associated proteins fewer epitope rich regions in lighter grey indicating a lower binding score. These observations are confirmed looking at the protein score and epitope density. The average protein score and average epitope density of the SARS-CoV-2 protein set (0.071, 0.108) are significantly higher than those of Influenza A (0.024, 0.044). Such *in-silico* comparisons can allow in future first insight into immune evasion due to longer pathogen-host co-evolution, DiscovEpi offers the *in-silico* basis for evaluating the involvement of a CD8+ T cell response in the defence against certain pathogens like here SARS-CoV-2. The tool significantly reduces time and workload doing epitope predictions for a set of proteins and allows a comprehensive global view on the potentially immunogenic sites of the proteins. Thus, DiscovEpi can serve as the first *in-silico* step in the search for new candidates for epitope-based drugs and vaccines against pathogens.

**word count: 1776**

**Funding information**

This work was supported by “Deutsche Forschungsgemeinschaft“(RTG 2719) and the Helmholtz Initialization and Networking Fund for Infection Research Greifswald (ZoonFlo: 2021).

References

Bröker, B.M. *et al.* (2016) The T Cell Response to Staphylococcus aureus. *Pathogens (Basel, Switzerland)*, **5,** 2016: 10.3390/pathogens5010031.

Da Silva Antunes, R. *et al.* (2023) An update on studies characterizing adaptive immune responses in SARS-CoV-2 infection and COVID-19 vaccination. *International immunology,* 2023: 10.1093/intimm/dxad014.

Damgaard, R.B. (2021) The ubiquitin system: from cell signalling to disease biology and new therapeutic opportunities. *Cell death and differentiation*, **28**, 423–426.

Dhanda, S.K. *et al.* (2019) IEDB-AR: immune epitope database-analysis resource in 2019. *Nucleic acids research*, **47**, W502-W506.

Jain, R. *et al.* (2021) Prediction of Epitope based Peptides for Vaccine Development from Complete Proteome of Novel Corona Virus (SARS-COV-2) Using Immunoinformatics. *International journal of peptide research and therapeutics*, **27**, 1729–1740.

Neefjes, J. *et al.* (2011) Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nature reviews. Immunology*, **11**, 823–836.

O'Donnell, T.J. *et al.* (2018) MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell systems*, **7**, 129-132.e4.

Pamer, E.G. *et al.* (1997) MHC class I antigen processing of Listeria monocytogenes proteins: implications for dominant and subdominant CTL responses. *Immunological reviews*, **158**, 129–136.

Parvizpour, S. *et al.* (2020) Epitope-based vaccine design: a comprehensive overview of bioinformatics approaches. *Drug discovery today*, **25**, 1034–1042.

Polack, F.P. *et al.* (2020) Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *The New England journal of medicine*, **383**, 2603–2615.

Reynisson, B. *et al.* (2020) NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic acids research*, **48**, W449-W454.

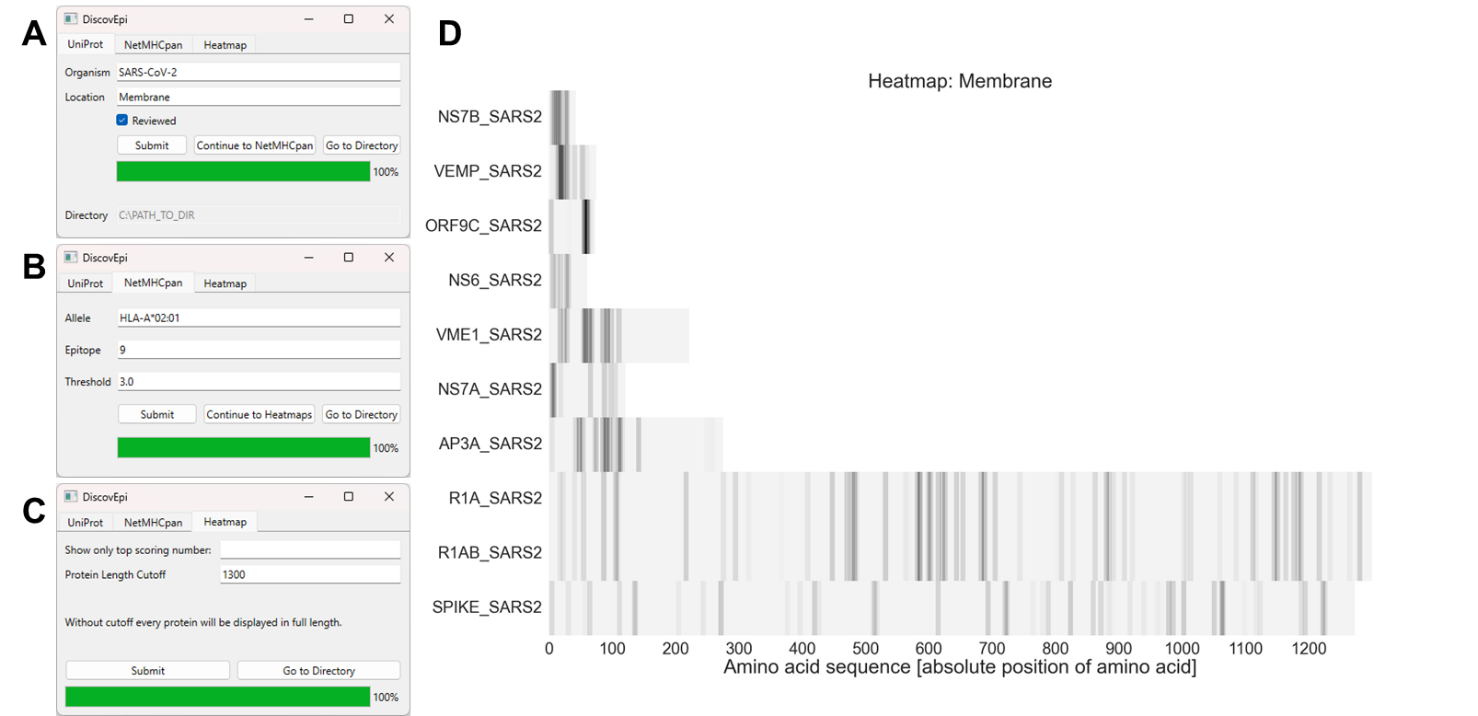
Russell, J.H. and Ley, T.J. (2002) Lymphocyte-mediated cytotoxicity. *Annual review of immunology*, **20**, 323–370.

Strehl, B. *et al.* (2005) Interferon-gamma, the functional plasticity of the ubiquitin-proteasome system, and MHC class I antigen processing. *Immunological reviews*, **207**, 19–30.

(2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic acids research*, **51**, D523-D531.

Waskom, M. (2021) seaborn: statistical data visualization. *JOSS*, **6**, 3021.

Wu, T. *et al.* (2019) Quantification of epitope abundance reveals the effect of direct and cross-presentation on influenza CTL responses. *Nature communications*, **10**, 2846.



**Figure 1: DiscovEpi schematic workflow and output.**

Workflow of DiscovEpi using (A) the protein sequences retrieval of the dataset and (B) the prediction parameter for a certain HLA allele, defined epitope length and NetMHCpan-score threshold. Visualization parameters (C) of the predicted epitopes with a field for the maximum number of top scoring proteins and maximum length for (D) the epitope map. The resulting heatmap is ordered by the protein-centric epitope density.

## **Supplemental Material**

## **Methods**

**DiscovEpi implementation**

The DiscovEpi algorithm is available on GitHub (https://github.com/cmahncke/DiscovEpi) containing XX packages for the usage under Windows, iOS, and Linux. DiscovEpi is implemented in Python (van Rossum, 2010) and the Qt framework using PySide for integration (The Qt Company, Qt for Python project; PySide, version 1.2.4. available at https://pypi.org/project/PySide). Seaborn visualization library is used to produce the epitope map in the form of a heatmap (Waskom, 2021). As interface between the UniProt database (UniProt: the Universal Protein Knowledgebase in 2023, 2023) and NetMHCpan (Reynisson *et al.*, 2020) we implemented REST APIs hosted by UniProt and IEDB (Dhanda *et al.*, 2019). NetMHCpan (version 4.1) is implemented in the python scripts and executables. NetMHCpan generates a score of predicted strength of peptide-MHC binding, which is based on a trained neuronal network. This neuronal network calculates the strength of binding based on amino acid properties, peptide-MHC interactions and experimentally measured binding affinities (Reynisson *et al.*, 2020). This score is compared to the distribution of scores in a large, maintained reference library, and NetMHCpan provides a percentile rank as a measure of the relative strength of the predicted peptide-MHC-I binding.

The file also contains an overview about the query parameters and retrieved sequences and is named “unp\_ORGANISM\_LOCATION.xlsx” where the bold letters are replaced by the respective input. After epitope prediction the retrieved epitopes with position and normalized binding score, the described protein scores, and meta-data (UniProt-ID, Protein name, UniProt-Link) are saved to a spreadsheet named “nmp\_ORGANISM\_LOCATION\_ALLELE.xlsx” in the previously specified directory. The file also contains data about the number of occurrences of each epitope and the query parameters.

**DiscovEpi epitope density and average binding score**

DiscovEpi allows the protein-centric search for well binding epitopes in whole proteomes based on the epitope density and average binding score of predicted epitopes. The binding score given by DiscovEpi is based on the percentile rank of each epitope as one of the output metrics from NetMHCpan. The percentile rank represents the relative binding affinity compared to a large reference group, including binding affinity scores for a diverse set of peptides and MHC molecules from experimental measurements and known MHC binding datasets. It is also used by DiscovEpi to discard peptides scoring worse than a specific percentile rank (default value = 3) , assuming these peptides are non-binders (Reynisson *et al.*, 2020). To compare the epitopes in the resulting limited set of high affinity binders the DiscovEpi epitope score is computed by normalizing the percentile rank to values between 0 and 1 (Formula 1).

Formula 1: Epitope density score. The percentile rank is normalized over the threshold (default value = 3). The difference between the threshold and percentile rank is divided by the threshold.

The epitope density score allows interpretation of protein sequences based on their density of potential epitopes. The epitope scores (Formula 1) are then used to compute the overall epitope score for the whole protein sequence by averaging over retrieved epitope scores and every possible epitope (Formula 2).

Formula 2: Protein density score. is the protein sequence length and  the epitope sequence length and the epitope scores of all retrieved epitopes for the respective protein sequence.

As the protein score does not differentiate between the presence of a few well binding or many weakly binding epitopes in a protein the epitope density is calculated as well (Formula 3). The epitope density of a protein defines the ratio of the number of predicted epitopes to the total number of possible oligopeptides of the same length in this protein.

Formula 3: Epitope density score. is the protein sequence length and the epitope sequence length and the epitope scores for all MHC class I epitopes retrieved for the respective protein. Here, the score itself is not important but the number of retrieved epitopes so that in combination with the protein density score (Formula 2) a holistic evaluation is possible.

**DiscovEpi visualization**

The epitope map is realized using a heatmap with amino acids positions on the x-axis and the proteins on the y-axis. So, each line of the heatmap describes one protein’s amino acid sequence. By default, the length of the x-axis matches the length of the longest protein in the set. Though, this value can be set individually as DiscovEpi takes the maximum length as input on the third GUI tab. The length of shorter proteins is illustrated as light grey background which represents protein but no epitope whereas white background represents no protein so automatically no epitopes. Numerically, the underlaying matrix contains the value 0.5 at each protein position and 0.0 when there is no protein. For each epitope of each protein the DiscovEpi scores are added to the default 0.5 in the respective line and amino acid sequence position (row). If there are overlapping epitopes the scores are added up. Visually, the intensity of the grey epitope marks is dependent on the calculated scores. So, the darker the epitope mark the higher the score. A high score here can describe few very probable epitopes or also many of less probability. Limiting the x-axis length increases resolution of shorter proteins since the shorter sequences are visualized using more horizontal space. Vertically, the resolution of the epitope map can be enhanced by setting a maximum number of proteins to be visualized on the map. This value can also be set on the third GUI tab. Especially bacterial protein sets can extend the resolution since the vertical height of the figure is fixed. The proteins shown on the map are ordered by the DiscovEpi protein score so that when limiting the number of proteins, the map still shows the most promising proteins. The resulting map is saved as PNG-file named “ORGANISM\_LOCATION\_heatmap.png” to the location specified on GUI tab one.