**Supplemental Material**

**1. Methods**

**DiscovEpi implementation**

The DiscovEpi algorithm is available on GitHub (https://github.com/cmahncke/DiscovEpi) containing packages for the usage under Windows and Linux. DiscovEpi is implemented in Python 3.10 (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 putative MHC-class I 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 class I molecules from experimental measurements and known MHC class I binding datasets. It is also used by DiscovEpi to discard peptides scoring below ? 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 essential 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 created using a heatmap with amino acid positions on the x-axis and the proteins on the y-axis. Each line of the heatmap describes the amino acid sequence of one protein. By default, the length of the x-axis matches the length of the longest protein in the set. However, this value can be set individually as DiscovEpi takes the maximum length as input selected on the third GUI tab (Figure 1C). 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 markings depends on the calculated scores i.e., the darker the epitope marking, the higher the score. A high score here can reflect few very probable epitopes or 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 (Figure 1C). 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 according to the DiscovEpi protein score so that even if the number is limited, 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 (Figure 1A).