DTU Health Tech
Department of Health and Technology

# Computationally predicting T-cell cross-reactivity induced autoimmunity with pathogenic proteome

Bachelor Thesis

# Abstract

Autoimmune diseases are caused by the immune system attacking the body's own tissues. There are several things that might cause this, including genetic, environmental, and immune system influences. One way by which autoimmune diseases develop is T-cell cross-reactivity, where a T-cell receptor activated by a pathogen peptide is also able to recognize self-peptides. This study investigates the potential for bacterial pathogens to cause autoimmune responses through T-cell cross-reactivity by comparing known autoimmune epitope sequences from the Immune Epitope Database (IEDB) against bacterial protein sequences from UniProt.

A dataset of 1,077 unique MHC class II-restricted human autoimmune epitopes was curated, and over 2 million pathogen proteins were obtained. All sub-epitopes of the autoimmune epitopes down to 9mers were created using a sliding window approach and matched using the Aho-Corasick algorithm. This process identified 7,188 potential epitope-pathogen cross-reactive matches. The majority of matches were from cytoplasmic bacterial proteins, which are less likely to be accessible for antigen presentation, suggesting that many of the matches may not contribute to autoimmune diseases. Matches were unevenly distributed across epitopes, with a few epitopes accounting for the majority of matches. Epitopes with most matches often came from proteins like heat shock proteins.

These findings indicate that pathogen-induced cross-reactivity is a rare event. Future models should include peptide similarity scoring and MHC/TCR binding sites to improve prediction accuracy.

# Acknowledgements

**Carolina Barra Quaglia**, Associate Professor, Department of Health Technology
Supervisor throughout the bachelor project. Many thanks for helping navigate through how to do a proper project and for all the feedback I got along the way.

**Alfred Ferrer Florensa**, PhD Student, National Food Institute
Cosupervisor throughout the bachelor project. Many thanks for all the extra time you put in, and for being very patient with me.

Both your help and contribution helped me to understand how to actually make, plan and structure a project. And your directions were invaluable.

# Introduction

Autoimmune disease arises when the body's immune system mistakenly targets its own cells, leading to chronic inflammation and damage to tissue. This loss of immune tolerance of the body's own cells is a result of recognition of self-antigens as foreign, leading to activation of autoreactive T-cells or the production of autoantibodies. Some of the most common and known autoimmune diseases include Type 1 diabetes (T1D), Rheumatoid arthritis (RA), Graves' disease and Hashimoto's thyroiditis [1].

Autoimmune diseases are often also classified into 2 main groups: organ specific, where the immune system targets specific organs, such as the thyroid in the case of Graves' diseases or hashimoto's thyroiditis. In systemic autoimmune diseases, multiple tissues or organs are attacked, as seen in Systemic lupus erythematosus or RA [2].

The pathology of autoimmune diseases contains many factors contributing to the development of the disease. For instance, T1D is most commonly caused by autoreactive antibodies recognizing beta-cell antigens as foreign, but there are some predispositions that increase the chance of developing these autoreactive antibodies. These include, but are not limited to, dysfunctional regulatory T-cells, specific HLA-alleles, gut microbiome, environmental factors and pathogenic infection [3]. These factors are not unique to T1D but can also influence the development of other autoimmune diseases as well.

While some autoimmune diseases are believed to be triggered by bacterial or viral infections [4], others have not been proven to have a pathogenic association. In these cases, the development of autoimmune diseases is often explained by a combination of  factors, including genetic predisposition, immune dysregulation, and environmental factors.

Several mechanisms have been proposed to explain how infections can trigger autoimmune diseases. Two of the most widely recognized are T-cell cross-reactivity and bystander activation.

Bystander activation occurs when inflammation from infection creates a cytokine-rich environment that promotes the activation of nearby autoreactive T cells, even without direct recognition of pathogen antigens [5].

T-cell cross-reactivity entails the T-cell receptor (TCR) of the T-cells recognizing multiple similar antigens. If a T-cell clone is activated by a pathogen peptide similar enough to a self-peptide, it can result in autoreactive T-cells. This process is shown in figure 1 [6]. Originally, it was thought that T-cells were specific to a single antigen, but numeric evidence was presented and showed if this was true, there should be more than ~10^15 circulating T-cells present to recognize the possible peptide combinations [7]. This far exceeds the ~10^12 that is circulating, of which ~10^8 expresses a naive phenotype of TCR that cannot undergo affinity maturation. Therefore, these T-cells must, from the start, express TCR's that can recognize some of these ~10^15 possible peptides [8]. The implication of this is that for the immune system to be efficient, T-cells must be cross-reactive.

This study focuses on T-cell cross-reactivity as a driver of autoimmune disease, looking into the extent to which pathogen-derived peptides may resemble known autoimmune epitopes.

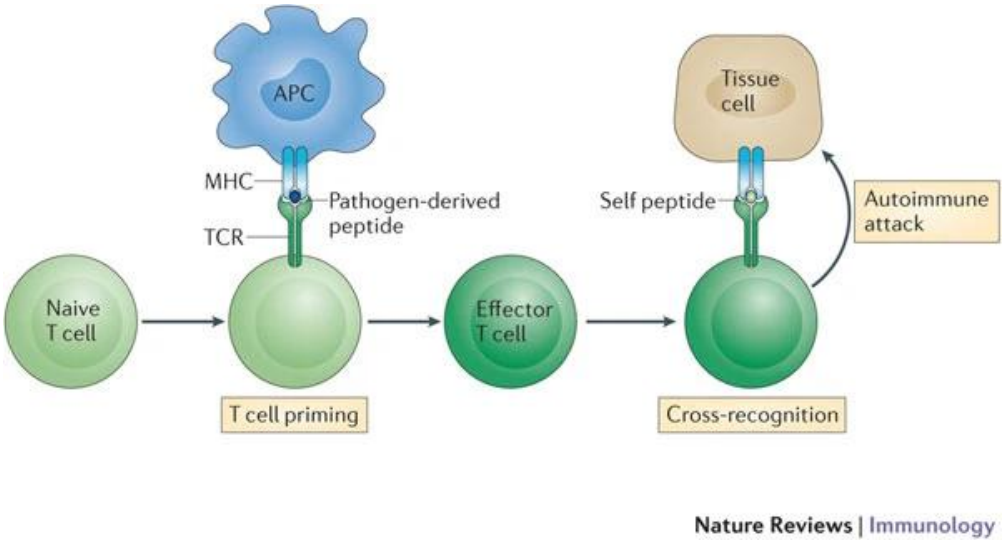


**Figure 1. Cross-reactivity of T-cell receptors leading to autoimmunity.** Naive T-cell activated by an APC carrying a pathogen derived peptide. The activated effector T-cell expresses a cross-reactive TCR that also recognizes a self-peptide, leading to an attack on self-cells. Figure taken from nature reviews, immunology.

Current methods for preventing or mitigating autoimmune diseases often involve immunosuppressive drugs. These drugs can reduce immune responses, but they can also promote the development of tolerogenic dendritic cells (tolDCs). TolDCs are characterized by low

expression of costimulatory molecules and by being able to induce T-cell anergy or cause regulatory T-cell differentiation. By increasing the number of tolDCs, immunosuppressive drugs help establish immune tolerance and alleviate autoimmune symptoms [9]. No apparent methods for the prevention of autoimmune diseases have been developed.

Understanding how autoimmune responses are initiated is crucial for understanding how pathogen-derived peptides can lead to autoimmunity. A key factor is how pathogen-derived peptides are processed and presented to T cells. Only a subset of pathogen proteins and corresponding peptides will actually be presented on MHC class II by Antigen Presenting Cells (APCs); this is determined by several factors such as the availability of the peptide region, its ability to bind to MHC class II and presence of T-cell repertoire able to recognize the peptide. This is also called the immunodominance of the peptides [10].

The subcellular location of proteins has been shown to correlate with the likelihood that they will be presented by APCs on MHC class II [11]. Immature dendritic cells (iDC's), often defined as professional antigen presenting cells, will often utilize macropinocytosis for surveillance, which involves taking in extracellular proteins. Macropinocytosis by the DC, in the absence of danger signals such as PAMPs (pathogen-associated molecular pattern), can help create self-tolerance, by iDC's presenting self-peptides to T-cells and inducing T cell anergy or differentiating the T-cell to T regulatory cells [12]. In the presence of danger signals the DC differentiates into activated dendritic cells (aDC). Activated dendritic cells utilize phagocytosis more, and they enter a presenting mode, by down-regulating the recycling of MHC class II and up-regulating the surface expression of MHC class II. Phagocytosis involves engulfing large proteins or sometimes entire pathogens, which enables dendritic cells to degrade extracellular, intracellular, and membrane proteins. These processes are shown in figure 2 [13]. The most readily available antigens for presentation will be the extracellular and membrane proteins.

Phagocytosis together with macropinocytosis, introduces a bias in antigen presentation towards extracellular and membrane proteins.

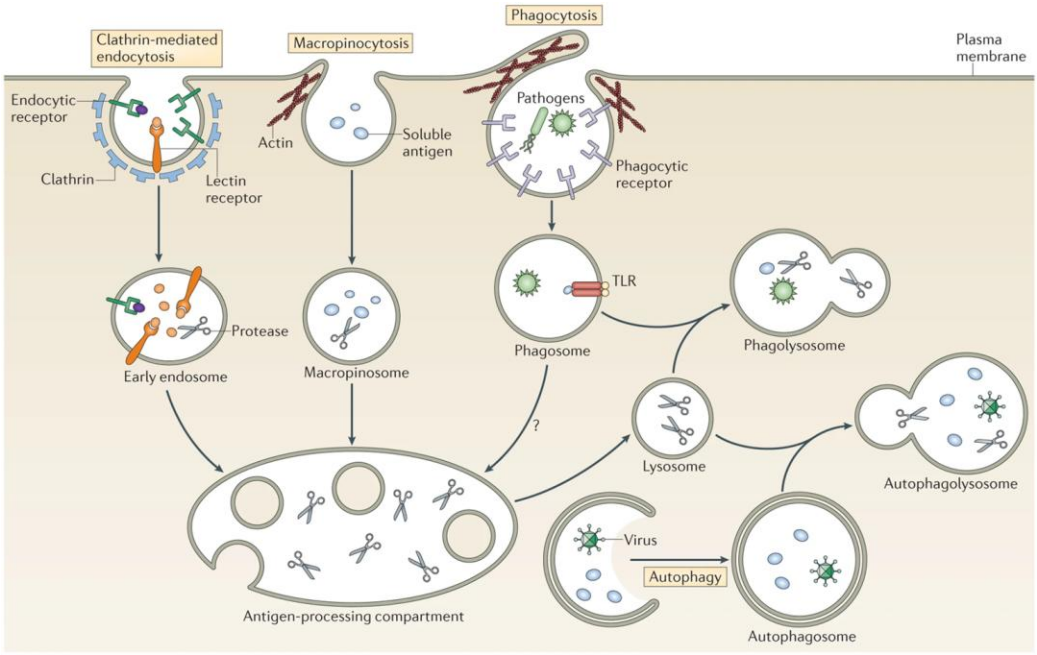


**Figure 2. Antigen uptake and processing by antigen-presenting cells (APCs)**. Antigens and pathogen can enter APC in different ways. Receptor-mediated endocytosis via clathrin-coated vesicles requiring antigen binding to a receptor on the APC. Macropinocytosis where the cell drinks from the surroundings by actin reformation. Phagocytosis where opsonized particles or pathogens bind to various receptors and enter the cell by actin reformation. The uptaken particles or pathogens are then degraded by proteases and enters an antigen-processing compartment where peptides for MHC II are generated. Figure taken from Nature Reviews, immunology.

Similarly, not all peptides within a protein are equally likely to be presented as epitopes. If an epitope is more exposed to solvent, it will be more accessible for proteases, thus more likely to be processed and presented by antigen-presenting cells (APCs) [14].

A model for predicting T-cell cross-reactivity to pathogen proteomes is yet to be developed and could be useful for identifying pathogens that potentially can cause autoimmune diseases.

Some autoimmune diseases have shown potential to be elicited by bacterial infection. This project will try to identify possible high-risk pathogen proteins involved in autoimmune diseases.

Specifically, it will investigate if known autoimmune epitopes, sourced from IEDB (Immune Epitope Database and Tools)[15], have identical sequences with pathogenic bacterial proteins, as

they could be candidates for T-cell cross-reactivity. It will look to discern patterns that might be associated with the epitope-pathogen protein match being involved in eliciting an autoimmune disease. To assess the immunogenic plausibility of these matches, we evaluate multiple biological features, including protein subcellular localization and solvent accessibility of the matched region.

This work will help identify possible high-risk pathogen proteins involved in autoimmune diseases and find patterns in the human epitopes involved in possible cross-reactivity.

# Methods

## Dataset Collection

Epitopes known to be associated with autoimmune diseases were retrieved from the Immune Epitope Database (IEDB) on February 15, 2025. Filters were applied to select linear epitopes derived from Homo sapiens, MHC class II restricted, associated with positive T-cell assays, and linked to autoimmune diseases. This filtering resulted in 4,890 assays covering 2,114 unique epitopes. An assay in IEDB refers to an experimental test that evaluates immune responses, such as T-cell activation or antibody binding, against a specific epitope.

Bacterial proteomes were downloaded from UniProt on February 15, 2025 [16], comprising 47,472 proteomes. The NCBI Pathogen Isolate Browser [17] was used to find organisms proven to be pathogenic. This contained 2.265.381 pathogens, which was downloaded on February 20, 2025. Filtering this by being a human host and presence of an assembly ID resulted in 970.042 pathogens. The filtered pathogen dataset was merged with the UniProt proteomes by assembly ID, resulting in 754 proteomes comprising 2,478,956 protein sequences.

## Epitope Data Preprocessing

The autoimmune epitope dataset was further filtered for redundancy by removing epitopes containing modified residues, filtering for epitope lengths between 12 and 25 amino acids, and

eliminating exact duplicates. For nested epitopes, meaning epitopes that overlapped, only the longest sequence was retained. This resulted in a final dataset of 1,077 unique epitopes across 1,077 assays.

## Pathogens Proteome Annotation

The Protein ID, Genus/Species, some strains, protein annotation and sequence was extracted from the proteomes. Since some proteomes did not contain strain information, an iterative search was implemented, by querying uniprot with the proteome ID and extracting the resulting strain information.

## Matching Epitopes to Pathogens

Potential cross-reactive matches were identified by comparing the autoimmune epitope sequences against the pathogenic protein sequences. To capture partial matches, a sliding window approach was used to generate all subsequences from each autoimmune epitope, going from full length down to 9-mers. This created many sub epitopes to compare to millions of pathogen proteins. To efficiently search for matches between the sub epitopes and the pathogen proteins an Aho-Corasick trie-based algorithm [18] was implemented. This created a tree structure from all the sub epitopes, and each pathogen protein was scanned against the tree. This resulted in many nested and duplicated results. To avoid redundant results, only the longest match between a given epitope and pathogen protein was kept. This process resulted in 7,188 unique epitope-pathogen-protein matches.

## Protein Localization Prediction

Predictions of protein subcellular localization were performed using DeepLoc 2.1 [19] for the IEDB epitope source proteins and DeepLoc Pro 1.0 [20] for the matched pathogenic proteins.

### Secondary Structure Prediction

Relative Solvent Accessibility (RSA) of the epitope source proteins and the matched pathogen proteins were predicted using NetSurfP 3.0 [21].

## Results

### Uniprot pathogenic proteomes

Bacterial proteomes were obtained from uniprot and cross referenced with NCBI pathogen isolate browser, to include only pathogenic bacteria.

The bacterial proteomes were not clustered to remove redundancy, as this was necessary to ensure the ability to trace back each match to the pathogen that caused the match. This approach might introduce imbalances, as some species may have more proteomes than others from the filtered uniprot data. To investigate potential proteome differences, the number of proteomes per species were counted, as species with more strains would contribute more proteomes.

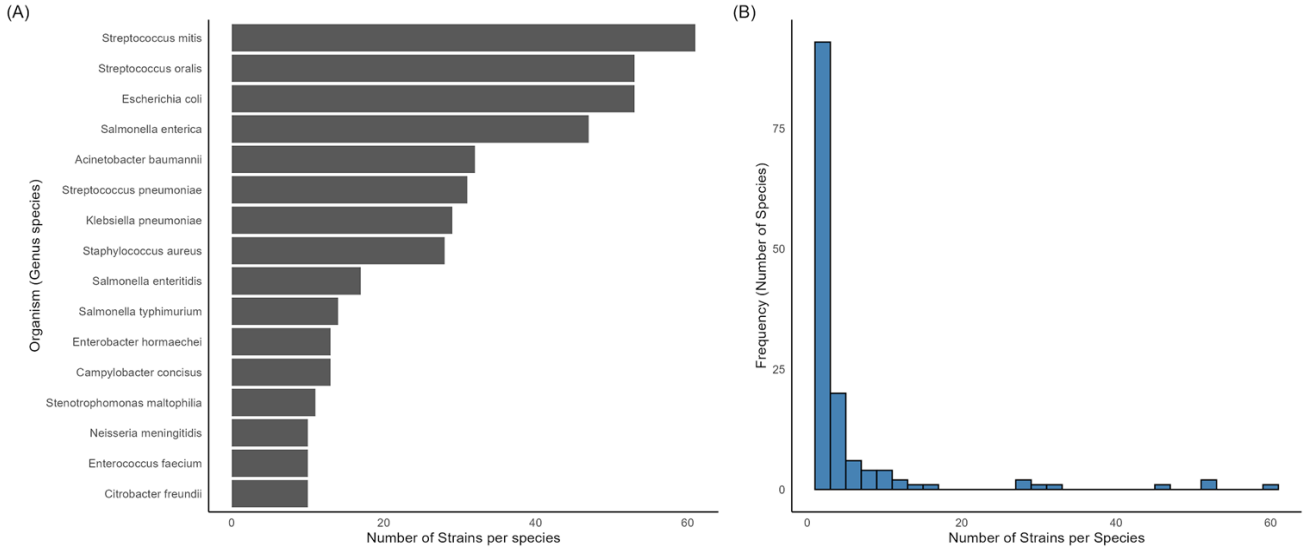


**Figure 3. Distribution of the number of strains per pathogen species in the UniProt dataset.** (A) The 15 most common pathogen species, ranked by the number of strains. (B) Histogram showing the distribution of strain counts across all pathogen species.

It is evident that most species are represented by only a few strains (Figure 3B), while a few species, such as *Streptococcus mitis*, *Streptococcus oralis*, and *Escherichia coli*, have up to 60 strains each (Figure 3A). These imbalances in strain numbers can strongly influence the number of matches identified per species.

However, the size of bacterial proteomes also varies considerably between species. Even a species with many strains may have a small proteome perhaps causing fewer matches. To investigate this, the total amino acid content of each species was quantified, by summing the amino acids across all its strains. This provides an estimate of each species' potential to generate matches.

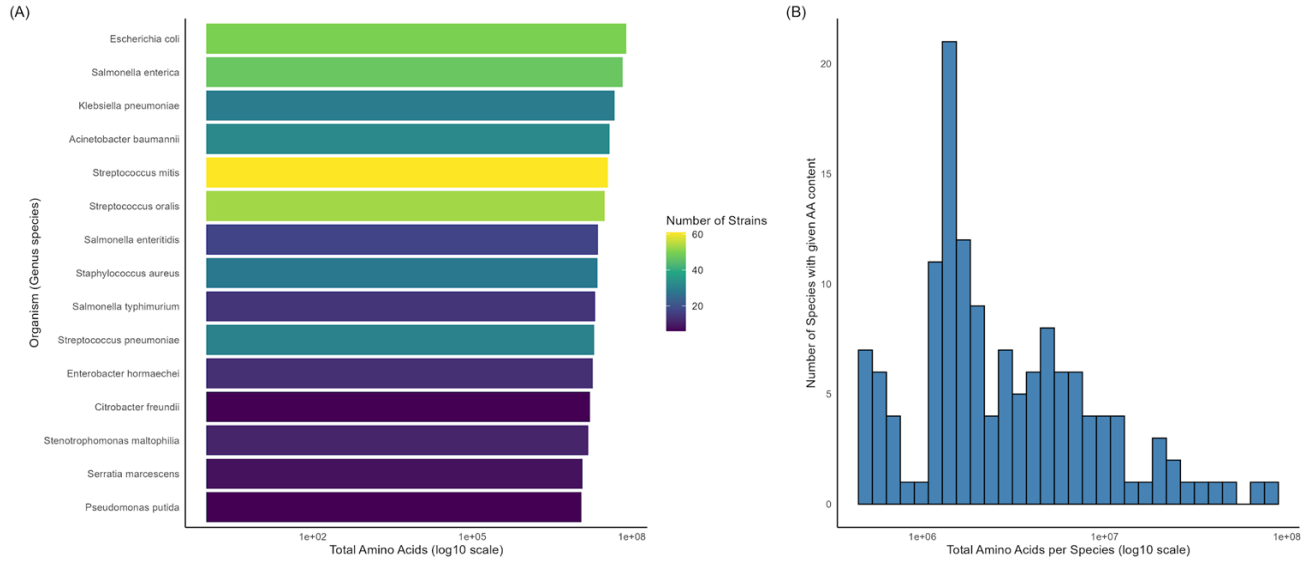


**Figure 4. Distribution of total amino acid content per pathogen species in the UniProt dataset.** (A) The 20 species with the highest total amino acid content, ranked by the combined amino acid length across all strains. (B) Histogram showing the distribution of total amino acid content across all species (log10 scale). Colour indicates the number of strains per species.

The distribution of amino acid content per species across all strains varies a lot (Figure 4B). We see that most species have around a hundred thousand to a million amino acids. However, some species stand out as outliers, with summed proteome sizes close to 10 million amino acids, such as *E. coli, S. enterica and K. pneumonia* (Figure 4A). Most of the species in the top 15 in amino acids were also in the top 15 for most strains. This could potentially be caused by the species with more strains being more extensively studied and thereby having more of its proteome covered.

## IEDB data

The autoimmune epitope dataset was filtered to remove modified residues, get only epitopes with lengths 12–25 amino acids, eliminate exact duplicates, and retain only the longest sequence from overlapping (nested) epitopes. This resulted in 1,077 unique epitopes across 1,077 assays.

Understanding the distribution of epitope source proteins in the IEDB dataset gives insight into which proteins are most commonly studied. If great imbalances are present, it could yield more false positive results for the more studied epitope source proteins, given they have more epitopes that can potentially match. To explore this, the number of unique epitopes for each epitope source protein was summed after the filtering of the dataset.

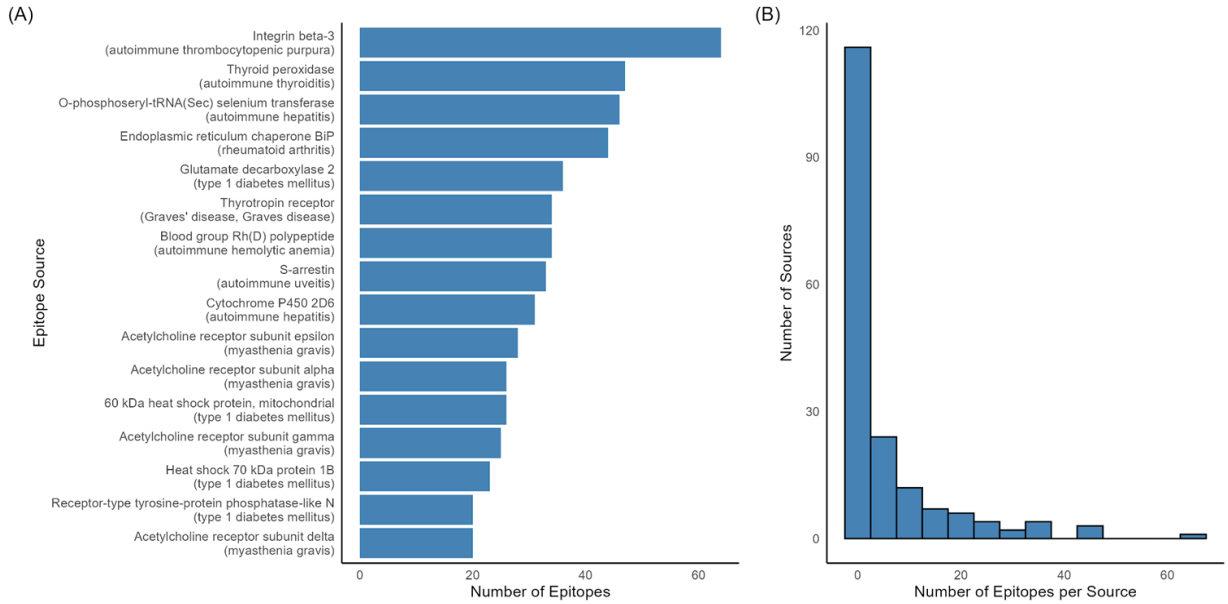


**Figure 5. Distribution of epitope counts per source protein in the IEDB dataset.** (A) The 15 source proteins with the highest number of epitopes and the disease they cause in parenthesis. (B) Histogram showing the distribution of epitope counts across all source proteins.

The distribution of epitopes per protein revealed that most proteins had only 1–5 epitopes studied (Figure 5B). A small number of proteins had most of the epitopes, with counts increasing exponentially. The top 15 proteins studied came from 8 different autoimmune diseases (Figure 5A) from a total of 33 autoimmune diseases present in the filtered IEDB dataset.

## Match data

Potential cross-reactive matches were identified by comparing autoimmune epitope sequences against pathogen protein sequences. To capture partial matches, all subsequences of each epitope down to 9-mers were generated. An Aho-Corasick trie-based algorithm enabled efficient searching of these sub-epitopes against the millions of pathogen proteins. To reduce redundancy, only the longest match between each epitope and pathogen protein was retained. This process resulted in 7,188 unique epitope–pathogen protein matches.

The matching process generated a large number of potential cross-reactive matches between autoimmune epitopes and pathogen proteins. To better understand how these matches were distributed, the amount of pathogen proteins matched to each epitope was analyzed. This analysis helps reveal whether a few epitopes dominate the matches or if matches are evenly distributed across many epitopes.

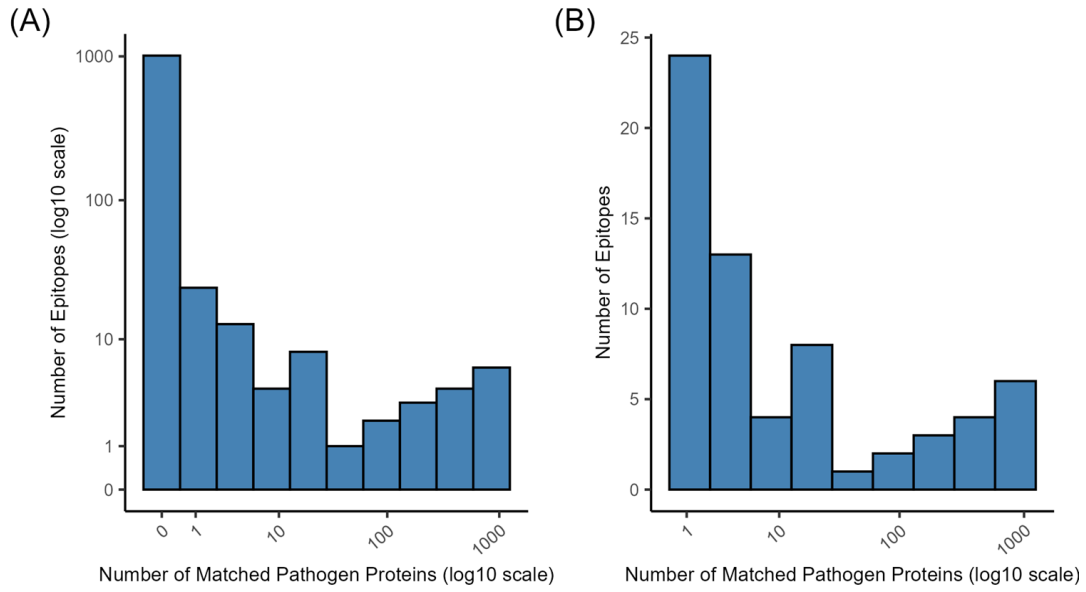


**Figure 6. Distribution of matched pathogen protein counts across autoimmune epitopes.**
(A) Histogram showing the distribution of matched pathogen protein counts across all epitopes.
(B) Histogram showing the distribution of matched pathogen protein counts for epitopes with at least one match. Both axes are shown on a log10 scale for clarity.

The vast majority of autoimmune epitopes did not get any matches with pathogen proteins (Figure 6A). Among the epitopes that did produce matches, most had only a few, while a small number of epitopes emerged as outliers, with some matching to close to a thousand pathogen proteins.

The species diversity per epitope was assessed to determine if these matches were associated with many species or primarily a few. To explore this, the joint distribution of match count and species diversity across epitopes was examined.

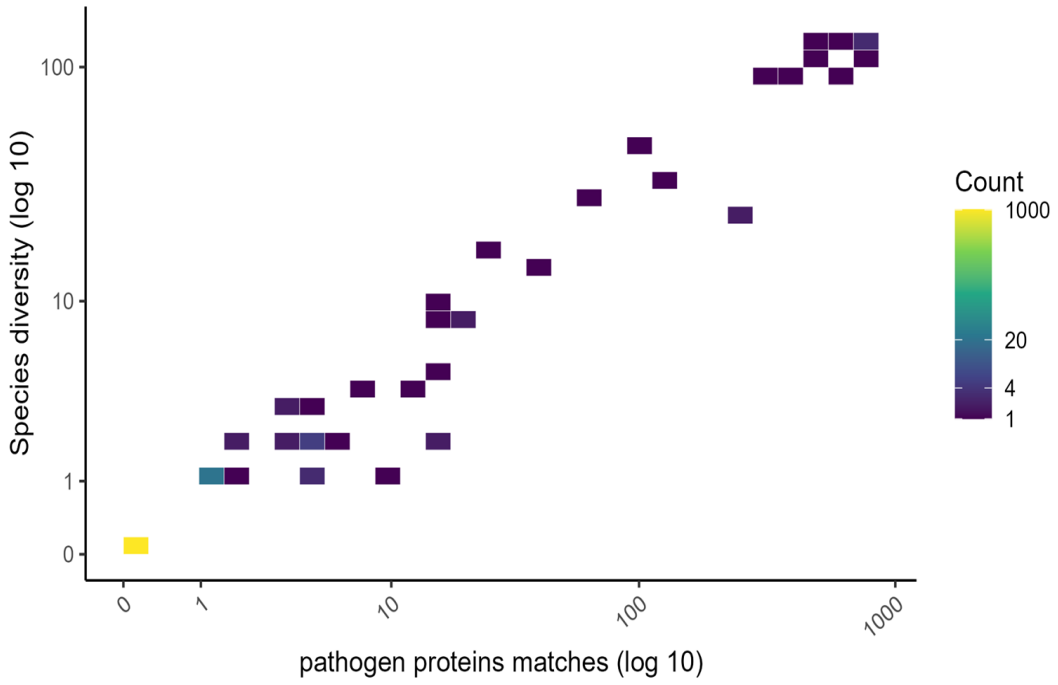


**Figure 7. Two-dimensional histogram showing the distribution of matched epitopes by number of matched pathogen proteins (x-axis) and species diversity (y-axis).** Each square represents a bin, with color indicating the number of epitopes falling into that combination. Axes are shown on a log10 scale.

As the number of matched pathogen proteins increases, the number of species involved also increases (Figure 7). This suggests that matches are not typically driven by a single species, instead epitopes with many matches tend to be associated with a broad diversity of species. For epitopes with nearly 1,000 matches, almost all species in the pathogen dataset were represented. The epitope with the most matches had matched 139 of the total 140 species present in the filtered pathogen data.

Although these matches were distributed across many species, it was necessary to investigate whether some species contributed more matches than others. Identifying species that matched to autoimmune epitopes more, could reveal whether certain pathogens are more prone to cross-reactivity. To explore this, the number of times each species appeared in the matched dataset was counted.

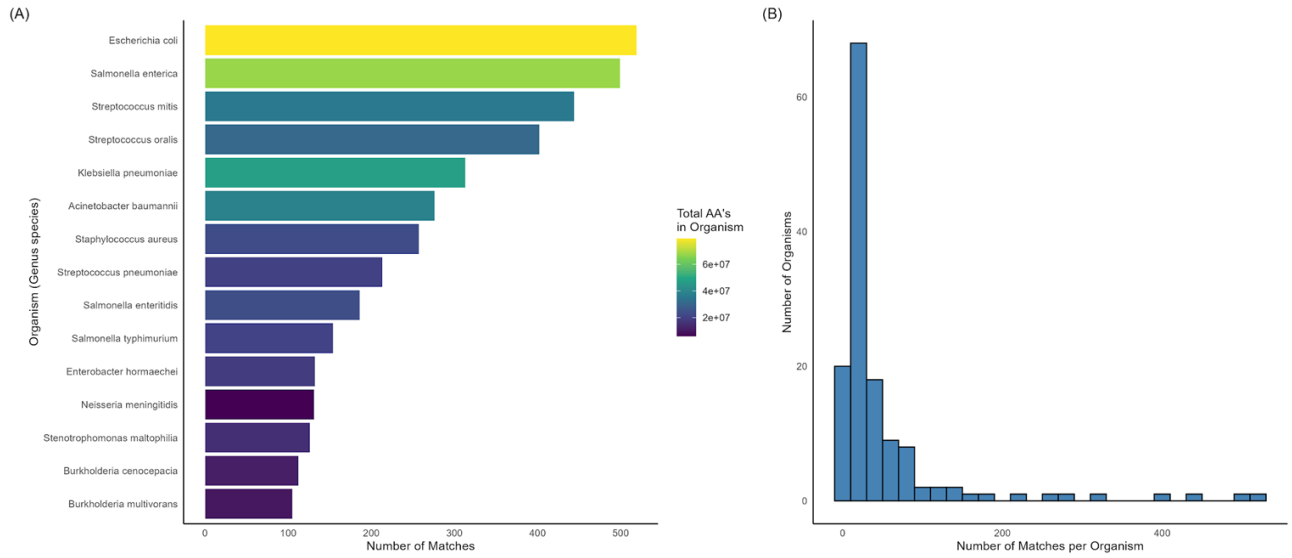


**Figure 8. Distribution of organism occurrences and total amino acid content in matched pathogen proteins.** (A) The 15 organisms with the highest number of occurrences in the matched pathogen dataset, colored by total amino acids per organism. (B) Histogram showing the distribution of organism occurrence counts across all species.

It was found that few species had disproportionately more matches than the rest, such as *E. coli*, *S. enterica* and *S. mitis* (Figure 8A). These were also some of the species that ranked highest in total amino acids content and number of proteomes (Figure 3 and 4). It was also found that most species had few occurrences in the matches, with around 50 occurrences each (Figure 8B).

Having asserted that certain species contribute more matches than others, the next step was to investigate which epitopes were causing these matches. This could help identify the associated epitope source proteins and autoimmune diseases that could be potentially more cross-reactive. To explore this, the distribution of matches per epitope was analyzed, along with its corresponding epitope source protein and autoimmune disease.

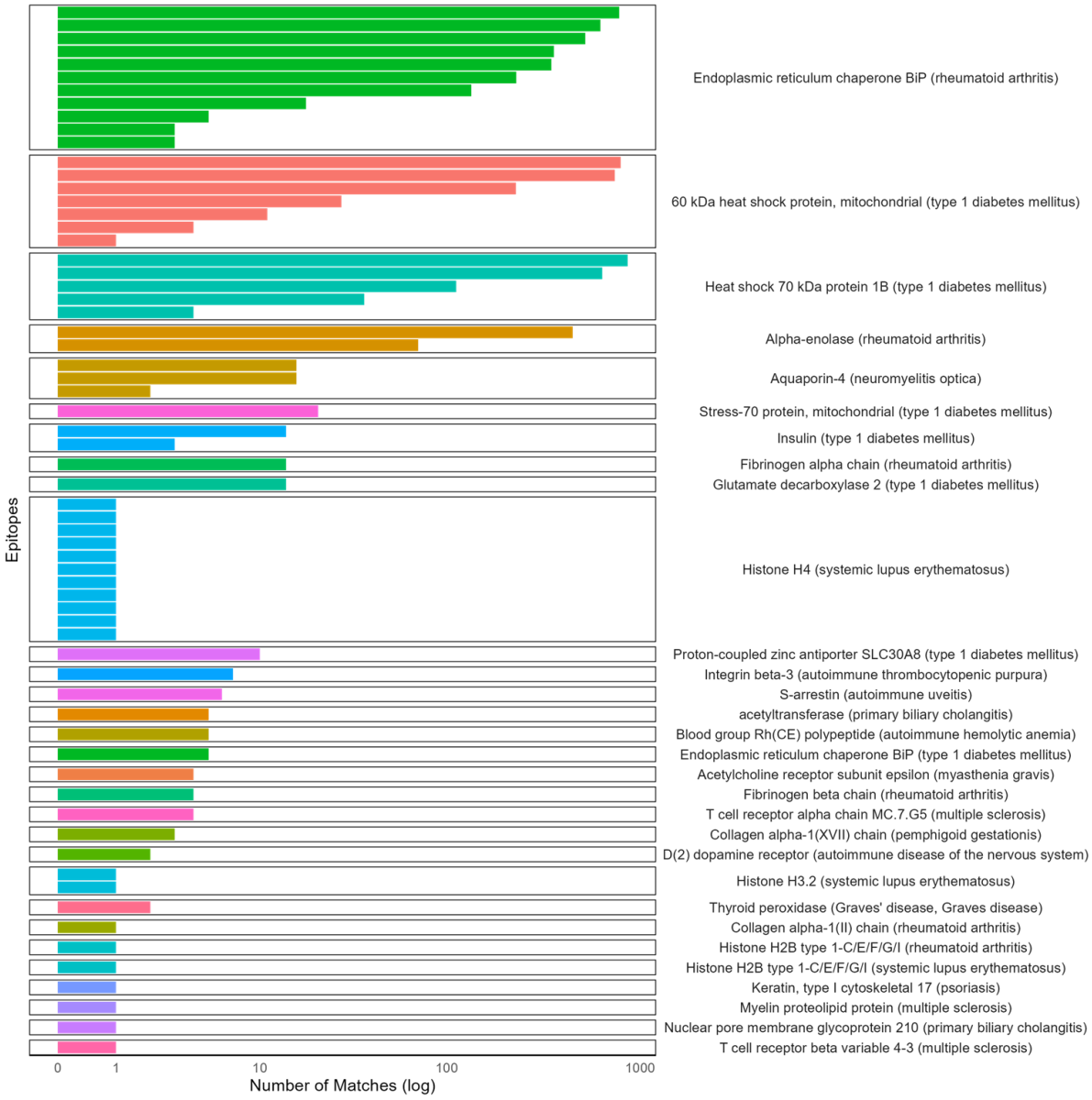


**Figure 9. Distribution of matched epitopes grouped by source protein and associated disease.** Each bar represents a single epitope, with bar height indicating the number of matched pathogen proteins. Facet panels are ordered by total match count per epitope source and annotated with the corresponding autoimmune disease.

Once again, we see that a few epitope source proteins account for the majority of matches. However, considerable variability exists between epitopes within the same source protein. This is clearly seen in the endoplasmic reticulum chaperone BiP and the 60 kDa heat shock protein, with some of their epitopes matching up to 1,000 pathogen proteins while others show only a few matches (Figure 9). The variability pattern is also seen in species diversity of epitopes (Figure 18, appendix).

This disproportionate number of matches for some epitope source proteins could potentially be explained by high sequence identity between the epitope source protein and its matched pathogen proteins. To determine this, the sequence identity between each epitope source protein and its matched pathogen proteins was analyzed by global alignment.

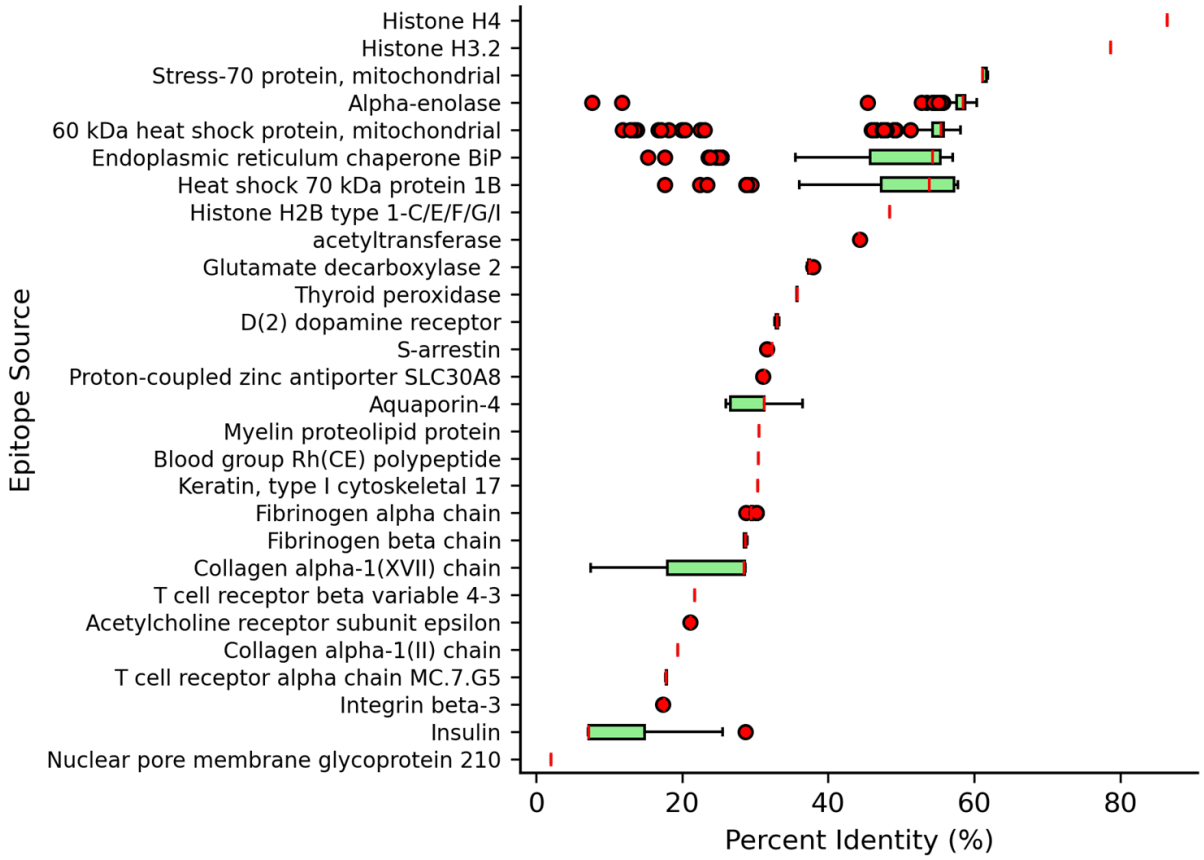


**Figure 10. Distribution of percent identity between pathogen proteins and the corresponding matched epitope source protein.** Each box represents the distribution of the percent identity of the pathogen proteins that matched to the given epitope source.

Epitope source proteins with a high number of matches (Figure 9) also tend to show higher median sequence identity with the matched pathogen proteins (Figure 10). A notable exception is the histone protein, which, despite having relatively few matches, exhibits near-complete sequence identity with the pathogen proteins it matches.

Since many of the high-match epitopes identified earlier were associated with a few specific diseases, the distribution of match counts per disease was examined. This analysis could help identify whether certain diseases are more susceptible to pathogen-induced cross-reactivity.

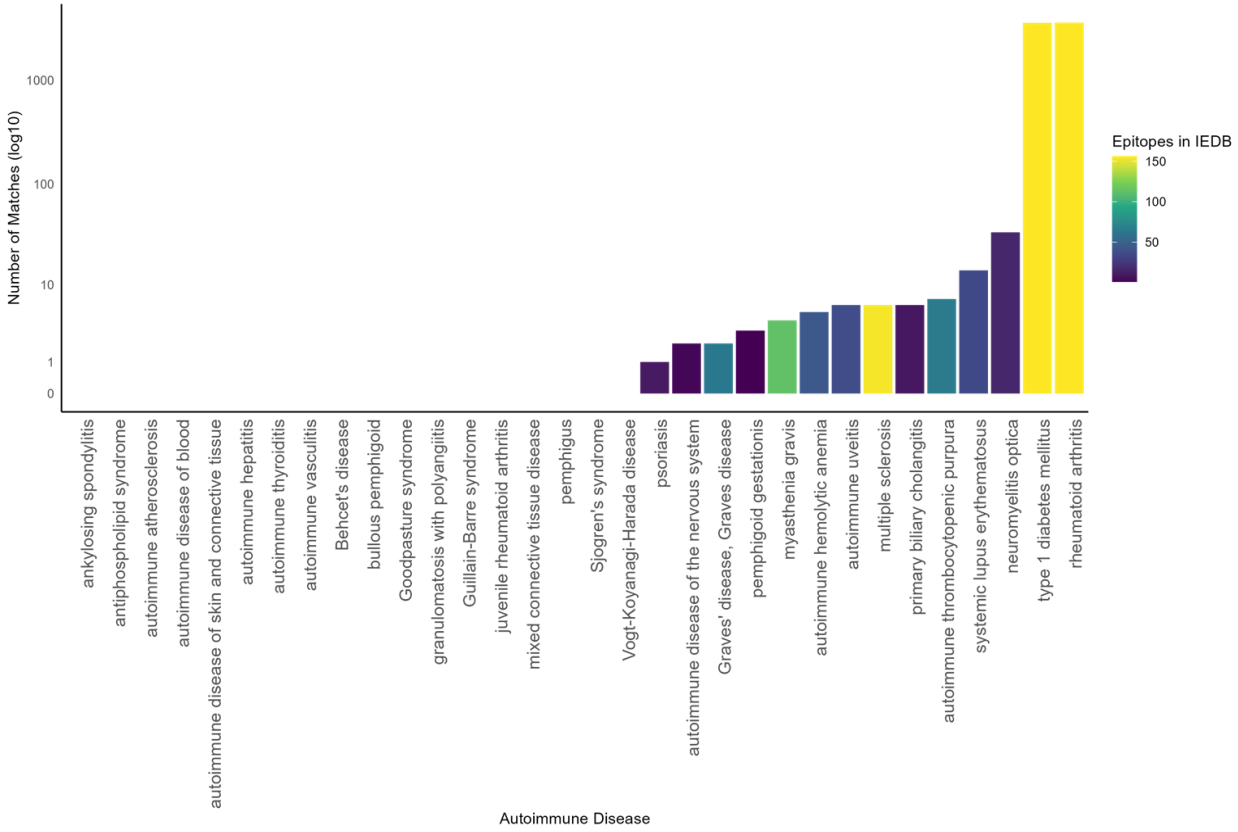


**Figure 11. Distribution of matched pathogen proteins by associated autoimmune disease.** Bar height indicates the number of matched pathogen proteins for each disease, while bar color shows the total number of epitopes originally associated with that disease in the IEDB database.

Most diseases showed little to no involvement in matches with pathogen proteins (Figure 11). Whereas rheumatoid arthritis and type 1 diabetes stood out, each accounting for thousands of matches. The remaining diseases were associated with only a few matches.

## Immunogenic analysis of matches

Given the large number of matches and the diverse set of pathogen proteins involved, the subcellular location of the matched pathogen proteins was assessed for their potential relevance

for cross-reactivity. To achieve this, all matched pathogen proteins were analyzed using DeepLoc Pro, and the predicted subcellular locations were quantified.

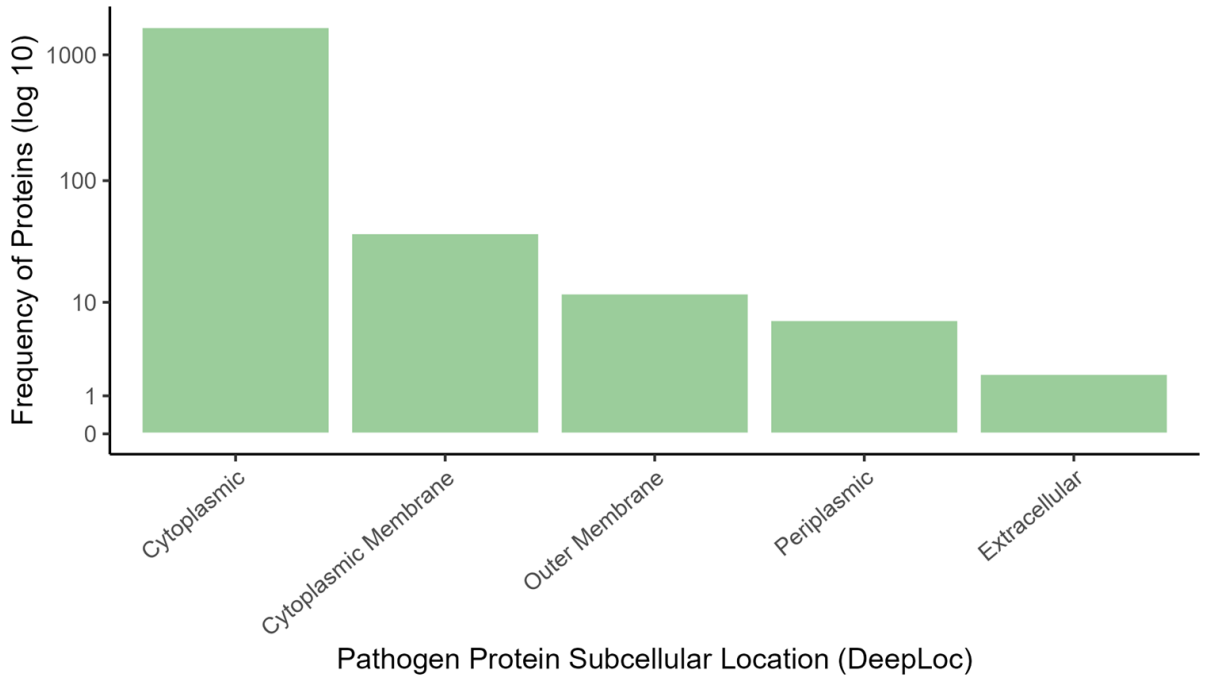


**Figure 12. Subcellular location distribution of matched pathogen proteins.** Distribution of the subcellular locations for the matched pathogen proteins. Predicted by deeploc pro. Y scale is log10 for visualization.

The majority of the matched pathogen proteins were predicted to be from the cytoplasm, with an exponentially decreasing amount being from the membrane and then extracellular (Figure 12).

To further analyze the subcellular location of the matched pathogen proteins, patterns were investigated between the number of matches an epitope received and the subcellular location of the matching pathogen proteins. For this, matches were grouped by epitope and subcellular location of the matched pathogen proteins, and these groups were visualized in relation to the total number of matches and pathogen species an epitope had.

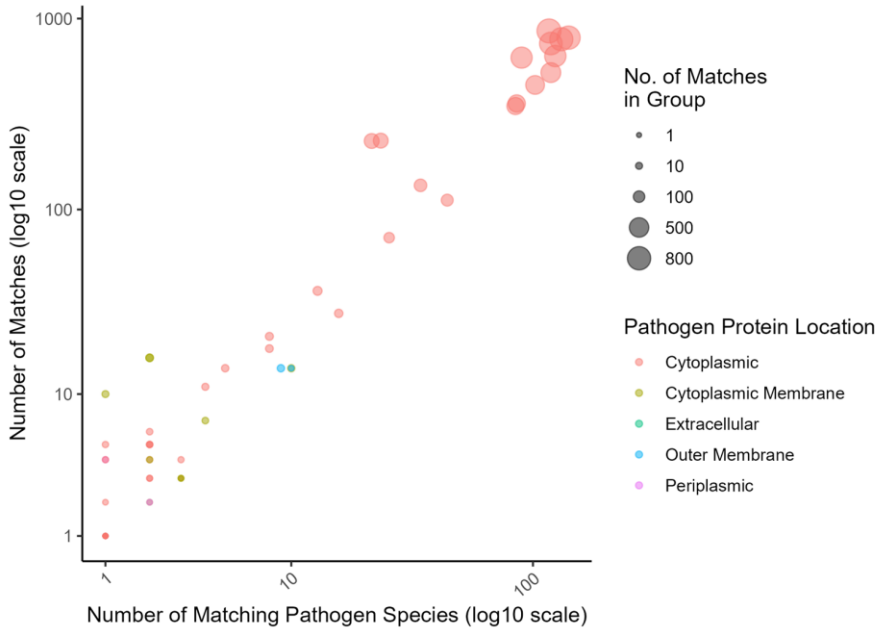


**Figure 13. Distribution of epitope matches by pathogen protein subcellular location.** Each point represents a group of pathogen proteins with the same subcellular location matched to a specific epitope. The x-axis shows the number of matching pathogen species per epitope (log10 scale), and the y-axis shows the total number of matched pathogen proteins (log10 scale). Point size indicates the number of pathogen proteins within each group, and color indicates the subcellular location of the matched proteins.

Epitopes with the highest number of matches only match with cytoplasmic pathogen proteins (Figure 13). In contrast, epitopes with fewer matches exhibit a greater diversity in the subcellular locations of the matching pathogen proteins.

While subcellular location provides important insights into potential protein accessibility, it does not capture the accessibility of individual peptides of the protein. Using NetsurfP, a prediction of how exposed each amino acid is on the pathogenic proteins and autoimmune epitopes was achieved. Using the relative solvent availability (rsa) value, predicted by NetsurfP, the matched peptide from the pathogen protein could be compared to the autoimmune epitope.

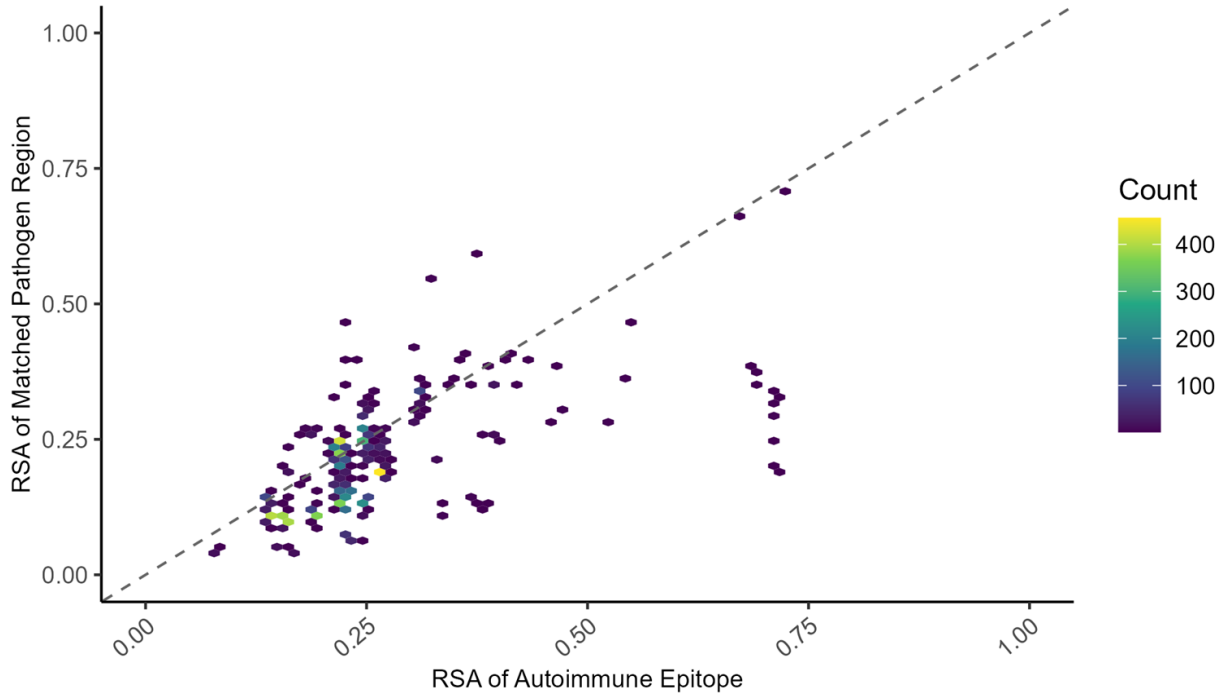


**Figure 14. Hexbin plot showing the joint distribution of relative solvent accessibility (RSA) values for matched pairs of autoimmune epitopes and pathogen peptide regions.** Each hexagon represents the number of matched pairs with a specific combination of RSA values. The dashed diagonal line indicates equal RSA values between autoimmune epitopes (x-axis) and pathogen peptides (y-axis). Hexagons above the line indicate cases where the pathogen peptide region has higher RSA than the autoimmune epitope, while hexagons below the line indicate the opposite.

In the majority of pairs, the autoimmune epitopes have higher RSA (Figure 14). The RSA range of both the immune epitopes and the matched pathogen peptides varies a lot, both having generally low rsa (Figure 17, appendix). Overall, no clear pattern emerges when comparing RSA values between pathogen peptides and autoimmune epitopes.

## Matched pathogens analysis

Given the wide array of matched pathogens, an investigation of which specific epitope source proteins and diseases these pathogens were associated with, was done to discern potential patterns of cross-reactivity. To focus the analysis on biologically plausible matches, the five epitope source proteins that exhibited the highest match counts (Figure 18, appendix) and the greatest species diversity were excluded. These epitopes came from cytoplasmic proteins, and

their widespread matching was considered unlikely, as such broad cross-reactivity does not align with the rarity of autoimmune diseases.

After removing these, the remaining epitope source proteins were analyzed to see which pathogens they matched with.

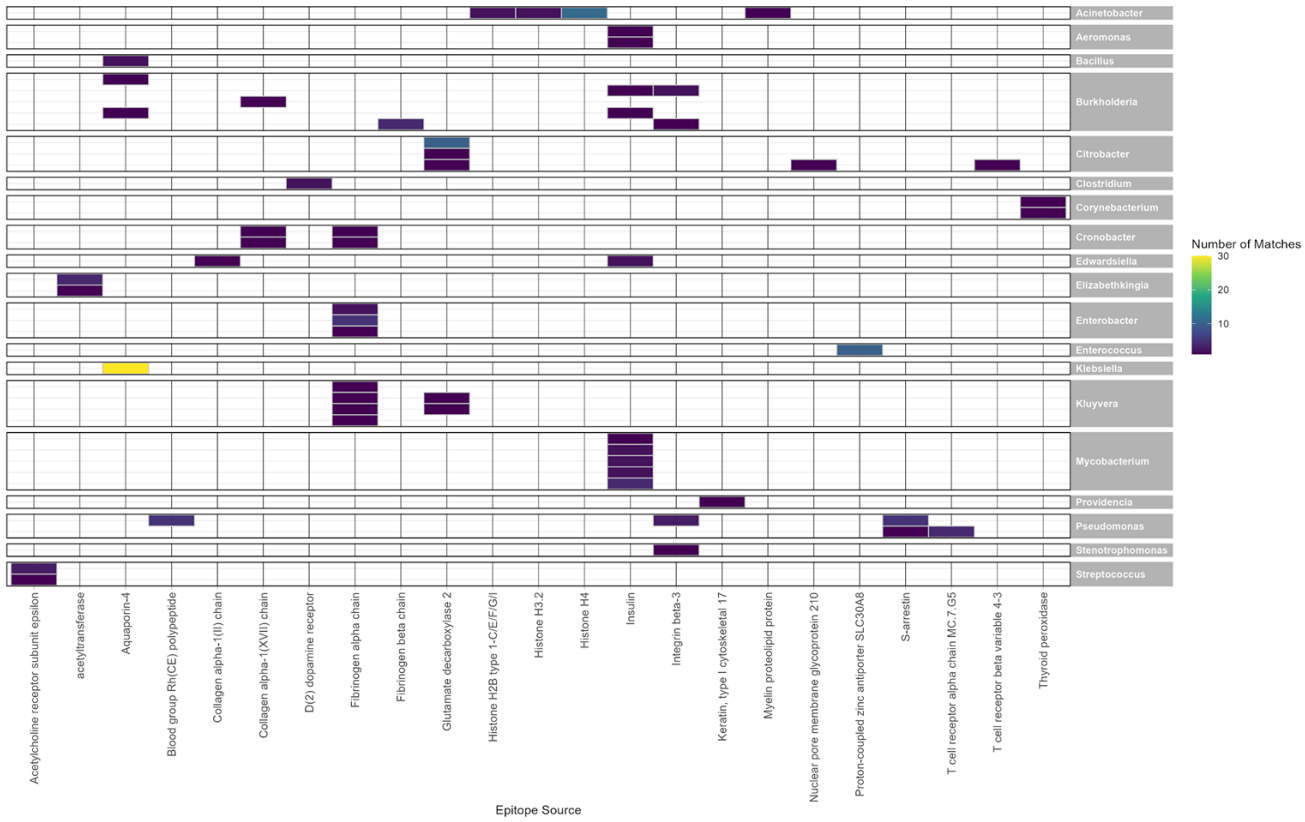


**Figure 15. Heatmap showing cross-reactivity potential between epitope source proteins and pathogen species.** Each tile represents the number of pathogen proteins from a specific species matched to a given epitope source protein. The plot is faceted by pathogen genus, with one row per species. Tile color indicates the total number of matches.

It is seen that in most cases the pathogens, by genus, match few epitope source proteins and that all species within a genus tend to match with few epitope source proteins (Figure 15). An exception to this is *Burkholderia,* which is matched with 5 different epitope source proteins

The pathogen association to a given epitope source protein gives us an overview of what disease they can potentially elicit, but not how immunogenic the instance is. To evaluate this, the subcellular location of the pathogen proteins for each epitope source protein was examined.

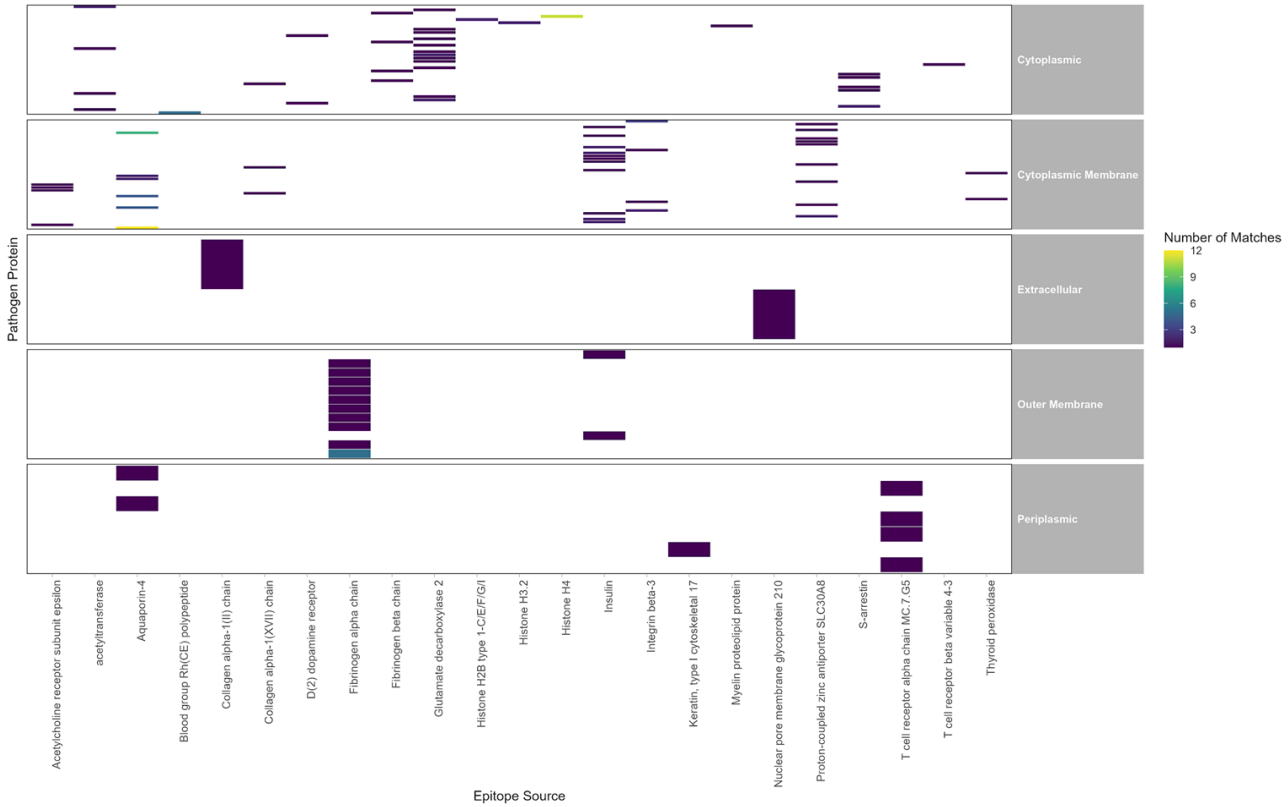


**Figure 16. Pathogen proteins and epitope sources by subcellular location.** Presence of matches between pathogen protein (y-axis) and epitope sources (x-axis). Faceted by the subcellular location of the pathogen proteins. Colored by the number of matches between pathogen protein and all epitopes from given epitope source.

Most of the matched pathogen proteins are located in the cytoplasm or the cytoplasmic membrane, and these matches are distributed across many epitope source proteins (Figure 16). Only a few matches were identified in the remaining subcellular locations. Interestingly, the only two matched extracellular proteins exclusively matched with a specific epitope source protein.

# Discussion

## Biological relevance of epitope matches

The study aimed to find potential cross-reactivity between known human autoimmune epitopes and pathogenic proteomes, with the goal of identifying possible bacterial infection that could led to the onset of autoimmune diseases. Through large scale comparison of millions of pathogenic proteins to around a thousand autoimmune epitopes, a few thousand matches was obtained, as seen in figure 6.

Autoimmune diseases affect roughly 5% of the population of the world and can be caused by multiple environmental and genetic factors [22]. The rarity of autoimmune diseases caused by pathogens, aligns with the observation that most of the autoimmune epitopes did not match with any pathogen proteins.

Few epitopes matched close to a thousand pathogen proteins as seen in figure 9. These include epitopes from the endoplasmic reticulum chaperone Bip, the 60/70 Kda heat shock protein and alpha-enolase. If a single epitope matched many pathogen proteins from a wide array of pathogens, it could suggest a higher likelihood of widespread cross-reactivity. However, these epitopes with many matches, matched purely with cytoplasmic proteins. Cytoplasmic proteins are generally less available for uptake by APCs. Since cross-reactivity does not seem to be a widespread phenomenon and the matched pathogen proteins are cytoplasmic, these epitopes and their matches were not considered likely cross-reactive events. To explore this further, the sequence identity between the epitope source proteins and the matched pathogen proteins was examined, as shown in Figure 10. This analysis revealed that epitope source proteins with high match counts also exhibited relatively high sequence identity to their matched pathogen proteins. Interestingly the histone epitope source protein, which had relatively few matches, was almost identical to its matched pathogen proteins. It was discovered that 92 of the 130 total species contained some histone related proteins, but only 1 species, *Acinetobacter lactucae,* was responsible for these histone matches. Given the unusually high sequence identity, it is possible that the histone in *A. lactucae* is coming from contamination introduced during proteome assembly. The histone match is therefore not considered a significant match.

These findings with high sequence identity raise important immunological considerations regarding the nature of tolerance and cross-reactivity. While conserved cytoplasmic proteins such as heat shock proteins often exhibit high sequence similarity across species, this alone does not necessarily imply that they trigger autoimmune responses. The repertoire of peptides used for negative selection in the thymus is diverse, but major providers for this repertoire include proteins most frequently involved with functional events [23]. Given that heat shock proteins are often constitutively expressed and evolutionary conserved [24] there is a high chance that tolerance against these is common. Therefore, finding the epitope sources showing characteristics of being less evolutionary conserved will be better candidates for potential cross-reactivity (Figure 10).

That said, there is nuance to this. All epitopes have been experimentally shown to ne autoreactive T cells, even the epitopes from conserved proteins. In such cases, the autoreactive T cells may arise due to context-specific mechanisms, such as self-antigens being presented to T cells during inflammation, infection, or other conditions that bypass central tolerance.

Since the extent of the project aims to find pathogenic proteins likely driving autoimmune diseases, looking at epitopes with high match count would then be biased towards evolutionary conserved sequences or implicitly assume that certain autoimmune diseases can be caused by a wide array of pathogens.

## Implications of pathogen protein subcellular location

From Figure 12, it is seen that most of the matches come from cytoplasmic proteins, with decreasing amounts observed as membrane and extracellular proteins. Cytoplasmic proteins are generally less available for uptake by APCs, making them less likely to be presented to T-cells; these matches would be less likely to cause autoimmune diseases. The scarcity of non-cytoplasmic subcellular location matches, speaks to the rare nature of T cell cross reactivity with pathogen proteins causing autoimmune diseases.

## Potential Pathogen Matches Associated with Autoimmune Diseases

After filtering the epitopes for not having relatively high match count and not matching cytoplasmic proteins, there were limited epitope-pathogen matches to investigate. The study focused on pathogen-epitope associations that may have relevance to specific autoimmune diseases. Factors considered included the known involvement of the pathogen species in the autoimmune disease and the subcellular location of the pathogen protein. The following sections summarize these key findings.

Top Candidate Pathogen Matches for Autoimmune Epitopes

| Autoimmune Disease | Epitope Source | Pathogen Species | Pathogen Protein Locations | Pathogen Proteome Sizes (AA) |
|---|---|---|---|---|
| Graves' disease, Graves disease | Thyroid peroxidase | Corynebacterium aurimucosum; Corynebacterium minutissimum | Cytoplasmic Membrane | 1.62e+06; 1.62e+06 |
| autoimmune thrombocytopenic purpura | Integrin beta-3 | Pseudomonas aeruginosa; Stenotrophomonas maltophilia; Burkholderia multivorans; Burkholderia cenocepacia | Cytoplasmic Membrane | 1.10e+07; 1.52e+07; 9.64e+06; 1.12e+07 |
| multiple sclerosis | T cell receptor alpha chain MC.7.G5 | Pseudomonas putida | Periplasmic | 1.13e+07 |
| myasthenia gravis | Acetylcholine receptor subunit epsilon | Streptococcus mitis; Streptococcus oralis | Cytoplasmic Membrane | 3.55e+07; 3.10e+07 |
| neuromyelitis optica | Aquaporin-4 | Klebsiella pneumoniae; Bacillus thuringiensis; Burkholderia ambifaria; Burkholderia latens | Cytoplasmic Membrane; Periplasmic | 4.74e+07; 1.55e+06; 2.07e+06; 2.01e+06 |
| pemphigoid gestationis | Collagen alpha-1(XVII) chain | Cronobacter malonaticus; Cronobacter sakazakii | Cytoplasmic Membrane | 1.30e+06; 1.26e+06 |
| primary biliary cholangitis | Nuclear pore membrane glycoprotein 210 | Citrobacter youngae | Extracellular | 2.96e+06 |
| psoriasis | Keratin, type I cytoskeletal 17 | Providencia stuartii | Periplasmic | 3.79e+06 |
| rheumatoid arthritis | Collagen alpha-1(II) chain | Edwardsiella tarda | Extracellular | 2.19e+06 |
| rheumatoid arthritis | Fibrinogen alpha chain | Kluyvera cryocrescens; Kluyvera intermedia; Enterobacter sichuanensis; Cronobacter malonaticus; Enterobacter roggenkampii; Kluyvera ascorbata; Enterobacter asburiae; Cronobacter sakazakii; Kluyvera georgiana | Outer Membrane | 1.56e+06; 1.65e+06; 1.53e+06; 1.30e+06; 7.83e+06; 1.55e+06; 4.66e+06; 1.26e+06; 1.46e+06 |
| type 1 diabetes mellitus | Insulin | Mycobacterium leprae; Mycobacterium lepromatosis; Mycobacterium haemophilum; Aeromonas hydrophila; Edwardsiella tarda; Aeromonas dhakensis; Mycobacterium tuberculosis; Burkholderia cenocepacia; Mycobacterium canetti; Burkholderia latens | Cytoplasmic Membrane; Outer Membrane | 5.75e+05; 5.12e+05; 1.21e+06; 1.60e+06; 2.19e+06; 1.43e+06; 5.21e+06; 1.12e+07; 1.29e+06; 2.01e+06 |
| type 1 diabetes mellitus | Proton-coupled zinc antiporter SLC30A8 | Enterococcus faecium | Cytoplasmic Membrane | 8.55e+06 |

**Table 1. Top candidate pathogen matches for autoimmune epitope.** Filtered by the matched pathogen protein not being cytoplasmic and by epitopes not matching to wide.

Autoimmune thrombocytopenic purpura (ITP) has matched with a few species, but all these species matched with a cation family protein. No confirmed linkage between the onset of ITP and bacterial infections has been reported, although ITP has been reported to follow viral infections [25].

The pathogenic proteins matched to ITP are all in the cytoplasmic membrane, which refers to the inner membrane in Gram-negative bacteria and in gram-positive the membrane beneath the cell wall. In both cases the bacteria would need to be phagocytosed to expose these proteins. This could explain why this particular match has not yet been confirmed as a driver of ITP and may represent a rare case.

Myasthenia gravis (MG), matched only with the *Streptococcus* genus. MG patients have shown alterations in their microbiota, compared to healthy. Specifically, *Streptococcus* and other bacteria showed higher abundance in patients with MG. In animal models it has also been shown that transferring microbiota from MG mice to healthy mice, resulted in reduced motor function of the healthy mice, which is a symptom of MG [26]. This has only been shown in a correlation aspect and no causation has been confirmed. A potential causal relationship, between the altered microbiome and the onset of MG, could be explained by the T-cell cross-reactivity between Acetylcholine receptor and the secretion protein from *Streptococcus,* but further research would have to confirm.

Type 1 diabetes (T1D), matched with *Edwardsiella tarda*. It has been shown that certain bacterial peptides are able to produce autoreactive T-cells against prepro-insulin peptides [27]. The species tested in this context was *Klebsiella oxytoca,* which is present in the dataset, but did not match with insulin epitopes. It is important to note that in their study they looked at MHC I autoimmune epitopes, whereas this study only contains MHC class II epitopes.

NMO (Neuromyelitis optica) pathogenesis, has many theories, including isoforms, hormones and genetics. Another is that it is initiated by viral or bacterial infection, in many cases infection preceded NMO pathogenesis [28]. It is mentioned that Mycobacterium and Mycoplasma sp. has homologs to the AQUA-4 protein, which is typically the immune target in NMO. In this study, the bacterial proteins that matched to the AQUA-4 epitopes included aquaporins and other general multipass membrane proteins. However, none of the previously suspected bacteria produced a match in these results.

In animal models, immunization with DnaK (a bacterial heat shock protein) from *Mycobacterium* produced T-cells able to cross react with human BiP proteins, suggesting

pathogenic infection can induce rheumatoid arthritis [29]. Many matches between bacterial DnaK and human endoplasmic reticulum BiP were observed in the results, with *Mycobacterium* being one of the many bacterias in the matches. However, given the ubiquity of heat shock proteins and their lower availability for APCs, it is likely bacterial infection is not the sole driver, and many other factors contribute to the development of rheumatoid arthritis.

## Conclusion

This study identified several potential cross-reactivity events between known human autoimmune epitopes and bacterial pathogen proteins. By using a sliding window approach to generate all possible sub-epitopes and using an Aho-Corasick algorithm for sequence matching, thousands of potential matches were found. The analysis revealed significant imbalances, with a few epitopes having the majority of matches, while most epitopes did not get any matches. This suggests that the matching needed more filtering and parameters to improve the relevance of the results.

One problem was that pathogen peptide sequences had to fit perfectly with the autoimmune sub-epitopes to match. This method made sure that the sequences were the same, but it assumed that T-cell cross-reactivity only happens between sequences that are exactly the same, which is not true. Matching down to 9mers could have also reduced the significance of the results, since longer matches could be more likely to cross-react. Increasing the match threshold would produce fewer results, so introducing substitution matrices like BLOSUM could allow mismatches and catch more results. Also, including the MHC-II anchor residue positions, could increase the relevance of the predicted matches. Including TCR binding positions could further refine the prediction by prioritizing peptide positions most likely to influence T-cell recognition.

## Literature


- [1] Hayter, S. M., & Cook, M. C. (2012). Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. *Autoimmunity Reviews*, *11*(10), 754–765. https://doi.org/10.1016/j.autrev.2012.02.001

- [2] Fridkis-Hareli, M. (2008). Immunogenetic mechanisms for the coexistence of organ-specific and systemic autoimmune diseases. *Journal of Autoimmune Diseases*, *5*, 1. https://doi.org/10.1186/1740-2557-5-1

- [3] Sundaresan, B., Shirafkan, F., Ripperger, K., & Rattay, K. (2023). The Role of Viral Infections in the Onset of Autoimmune Diseases. *Viruses*, *15*(3). https://doi.org/10.3390/v15030782

- [4] Getts, D. R., Spiteri, A., King, N. J. C., & Miller, S. D. (2020). Microbial Infection as a Trigger of T-Cell Autoimmunity. In *The Autoimmune Diseases* (pp. 363–374). Elsevier. https://doi.org/10.1016/B978-0-12-812102-3.00021-X

- [5] Pacheco, Y., Acosta-Ampudia, Y., Monsalve, D. M., Chang, C., Gershwin, M. E., & Anaya, J.-M. (2019). Bystander activation and autoimmunity. *Journal of Autoimmunity*, *103*, 102301. https://doi.org/10.1016/j.jaut.2019.06.012

- [6] Su, L. F., Kidd, B. A., Han, A., Kotzin, J. J., & Davis, M. M. (2013). Virus-specific CD4+ memory-phenotype T cells are abundant in unexposed adults. *Nature Reviews Immunology*, 13(6), 451–460. https://doi.org/10.1038/nri3279

- [7] Tagliamonte, M., Cavalluzzo, B., Mauriello, A., Ragone, C., Buonaguro, F. M., Tornesello, M. L., & Buonaguro, L. (2023). Molecular mimicry and cancer vaccine development. *Molecular Cancer*, *22*(1), 75. https://doi.org/10.1186/s12943-023-01776-0

- [8] Arstila, T. P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., & Kourilsky, P. (1999). A Direct Estimate of the Human αβ T Cell Receptor Diversity. *Science*, *286*(5441), 958–961. https://doi.org/10.1126/science.286.5441.958

- [9] Domogalla, M. P., Rostan, P. v., Raker, V. K., & Steinbrink, K. (2017). Tolerance through Education: How Tolerogenic Dendritic Cells Shape Immunity. *Frontiers in Immunology*, *8*. https://doi.org/10.3389/fimmu.2017.01764

- [10] Sadegh-Nasseri, S., & Kim, A. (2019). Selection of immunodominant epitopes during antigen processing is hierarchical. *Molecular Immunology*, *113*, 115–119. https://doi.org/10.1016/j.molimm.2018.08.011

- [11] Castro, A., Kaabinejadian, S., Yari, H., Hildebrand, W., Zanetti, M., & Carter, H. (2022). Subcellular location of source proteins improves prediction of neoantigens for immunotherapy. *The EMBO Journal*, *41*(24), e111071. https://doi.org/10.15252/embj.2022111071

- [12] Fucikova, J., Palova-Jelinkova, L., Bartunkova, J., & Spisek, R. (2019). Induction of Tolerance and Immunity by Dendritic Cells: Mechanisms and Clinical Applications. *Frontiers in Immunology*, *10*. https://doi.org/10.3389/fimmu.2019.02393

- [13] Roche, P. A., & Furuta, K. (2015). The ins and outs of MHC class II-mediated antigen processing and presentation. *Nature Reviews. Immunology*, *15*(4), 203–216. https://doi.org/10.1038/nri3818

- [14] Mahrus, S., Trinidad, J. C., Barkan, D. T., Sali, A., Burlingame, A. L., & Wells, J. A. (2008). Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell*, *134*(5), 866–876. https://doi.org/10.1016/j.cell.2008.08.012

- [15] Vita R, Blazeska N, Marrama D; IEDB Curation Team Members; Duesing S, Bennett J, Greenbaum J, De Almeida Mendes M, Mahita J, Wheeler DK, Cantrell JR, Overton JA, Natale DA, Sette A, Peters B. The Immune Epitope Database (IEDB): 2024 update. Nucleic Acids Res. 2025 Jan 6;53(D1):D436-D443. doi: 10.1093/nar/gkae1092. PMID: 39558162; PMCID: PMC11701597

- [16] The UniProt Consortium. *UniProt: the Universal Protein Knowledgebase in 2025*. Nucleic Acids Research. 2025;53:D609–D617. doi:10.1093/nar/gkad1049

- [17] National Center for Biotechnology Information. (n.d.). *NCBI Pathogen Detection Isolates Browser*. Retrieved from https://www.ncbi.nlm.nih.gov/pathogens/isolates/

- [18] Aho, A. v., & Corasick, M. J. (1975). Efficient string matching. *Communications of the ACM*, *18*(6), 333–340. https://doi.org/10.1145/360825.360855

- [19] Ødum, M. T., Teufel, F., Thumuluri, V., Almagro Armenteros, J. J., Johansen, A. R., Winther, O., & Nielsen, H. (2024). DeepLoc 2.1: multi-label membrane protein type prediction using protein language models. *Nucleic Acids Research*, *52*(W1), W215–W220. https://doi.org/10.1093/nar/gkae237

- [20] Moreno, J., Nielsen, H., Winther, O., & Teufel, F. (2024). Predicting the subcellular location of prokaryotic proteins with DeepLocPro. *Bioinformatics*, *40*(12). https://doi.org/10.1093/bioinformatics/btae677

- [21] Høie, M. H., Kiehl, E. N., Petersen, B., Nielsen, M., Winther, O., Nielsen, H., Hallgren, J., & Marcatili, P. (2022). NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Research*, *50*(W1), W510–W515. https://doi.org/10.1093/nar/gkac439

- [22] Shapira, Y., Agmon-Levin, N., & Shoenfeld, Y. (2010). Defining and analyzing geoepidemiology and human autoimmunity. *Journal of Autoimmunity*, *34*(3), J168–J177. https://doi.org/10.1016/j.jaut.2009.11.018

- [23] Collado, J. A., Guitart, C., Ciudad, M. T., Alvarez, I., & Jaraquemada, D. (2013). The Repertoires of Peptides Presented by MHC-II in the Thymus and in Peripheral

Tissue: A Clue for Autoimmunity? *Frontiers in Immunology*, *4*, 442. https://doi.org/10.3389/fimmu.2013.00442

- [24] Lanneau, D., Brunet, M., Frisan, E., Solary, E., Fontenay, M., & Garrido, C. (2008). Heat shock proteins: essential proteins for apoptosis regulation. *Journal of Cellular and Molecular Medicine*, *12*(3), 743–761. https://doi.org/10.1111/j.1582-4934.2008.00273.x

- [25] Pietras, N. M., Gupta, N., Justiz Vaillant, A. A., et al. (2025). Immune thrombocytopenia. *StatPearls*. Treasure Island (FL): StatPearls Publishing. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK562282/

- [26] Su, T., Yin, X., Ren, J., Lang, Y., Zhang, W., & Cui, L. (2023). Causal relationship between gut microbiota and myasthenia gravis: a bidirectional mendelian randomization study. *Cell & Bioscience*, *13*(1), 204. https://doi.org/10.1186/s13578-023-01163-8

- [27] Dolton, G., Bulek, A., Wall, A., Thomas, H., Hopkins, J. R., Rius, C., Galloway, S. A. E., Whalley, T., Tan, L. R., Morin, T., Omidvar, N., Fuller, A., Topley, K., Hasan, M. S., Jain, S., D'Souza, N., Hodges-Hoyland, T., Spiller, O. B., Kronenberg-Versteeg, D., … Sewell, A. K. (2024). HLA A*24:02–restricted T cell receptors cross-recognize bacterial and preproinsulin peptides in type 1 diabetes. *Journal of Clinical Investigation*, *134*(18). https://doi.org/10.1172/JCI164535

- [28] Graber, D. J., Levy, M., Kerr, D., & Wade, W. F. (2008). Neuromyelitis optica pathogenesis and aquaporin 4. *Journal of Neuroinflammation*, *5*(1), 22. https://doi.org/10.1186/1742-2094-5-22

- [29] Shoda, H., Hanata, N., Sumitomo, S., Okamura, T., Fujio, K., & Yamamoto, K. (2016). Immune responses to Mycobacterial heat shock protein 70 accompany self-reactivity to human BiP in rheumatoid arthritis. *Scientific Reports*, *6*(1), 22486. https://doi.org/10.1038/srep22486

- [30] https://github.com/luciostevns/Bachelor_project

# Appendix

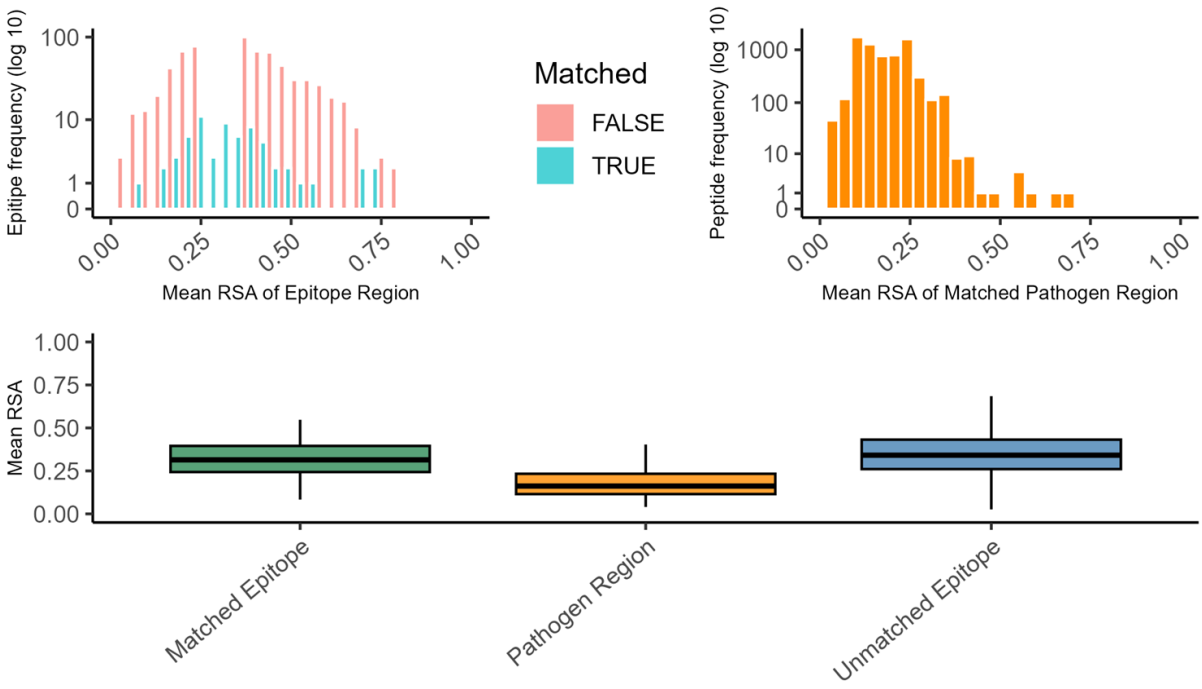


**Figure 17. Distributions of mean rsa values per peptides.** Top left plot shows mean rsa values for the autoimmune epitopes separated by if they matched to pathogenic proteins or not. Top right plot shows mean rsa values for the matched peptide region within the pathogen proteins. Bottom plot shows boxplot distribution of the matched epitopes, matched pathogen protein region and unmatched epitopes.

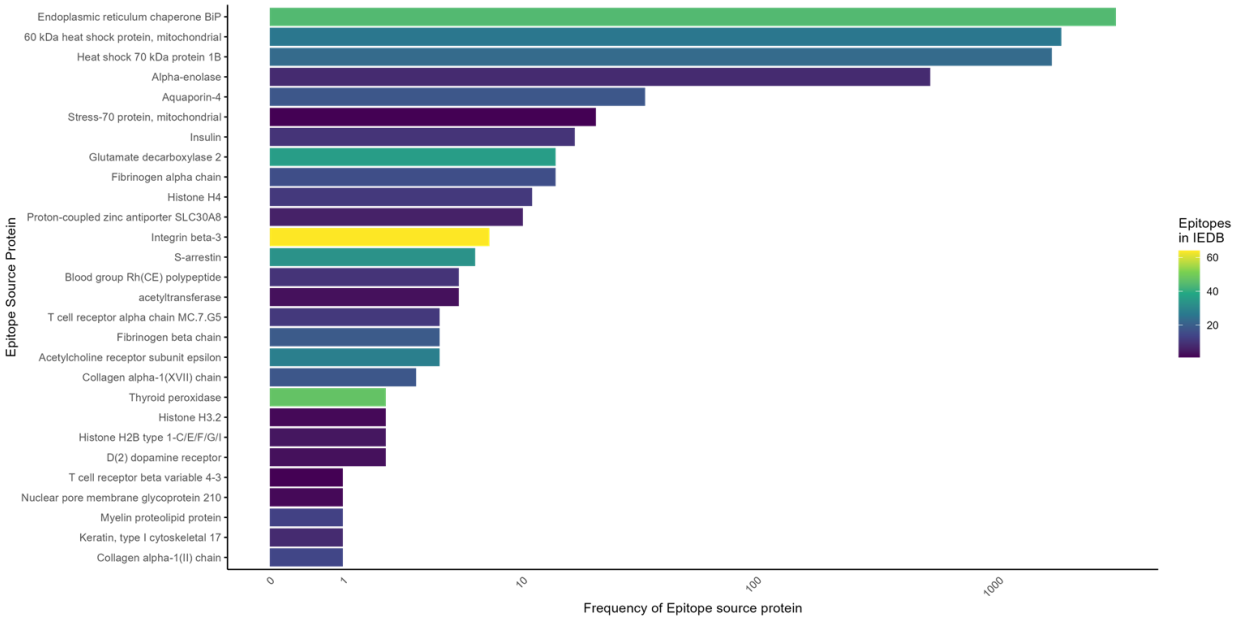


**Figure 18. Frequency of epitope source protein in the match data.** Colored by the original number of epitopes per protein source from IEDB data.

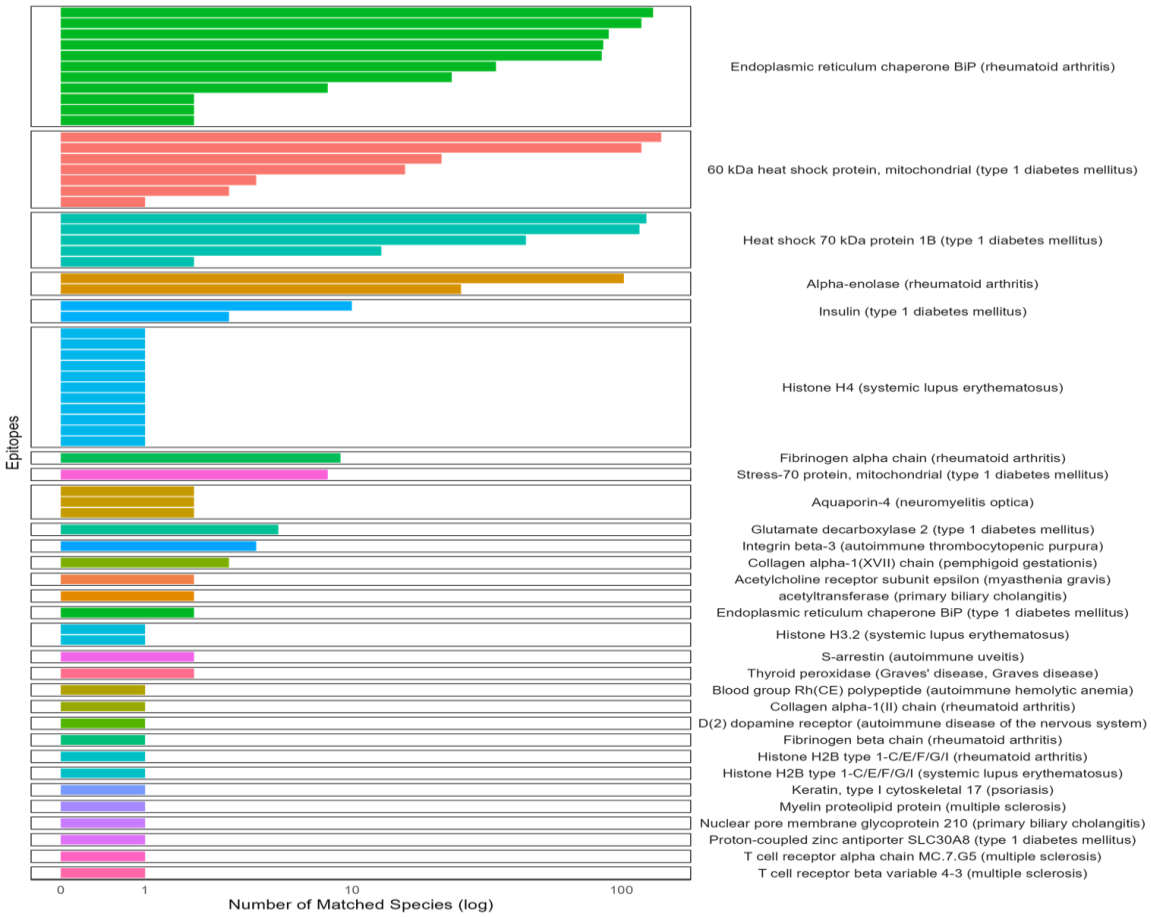


**Figure 19. Distribution of species diversity per matched epitope, grouped by source protein and associated disease.**
Each bar represents a single epitope, with bar height indicating the number of matched pathogen species. Facet panels are ordered by total species count per epitope source and annotated with the corresponding autoimmune disease.

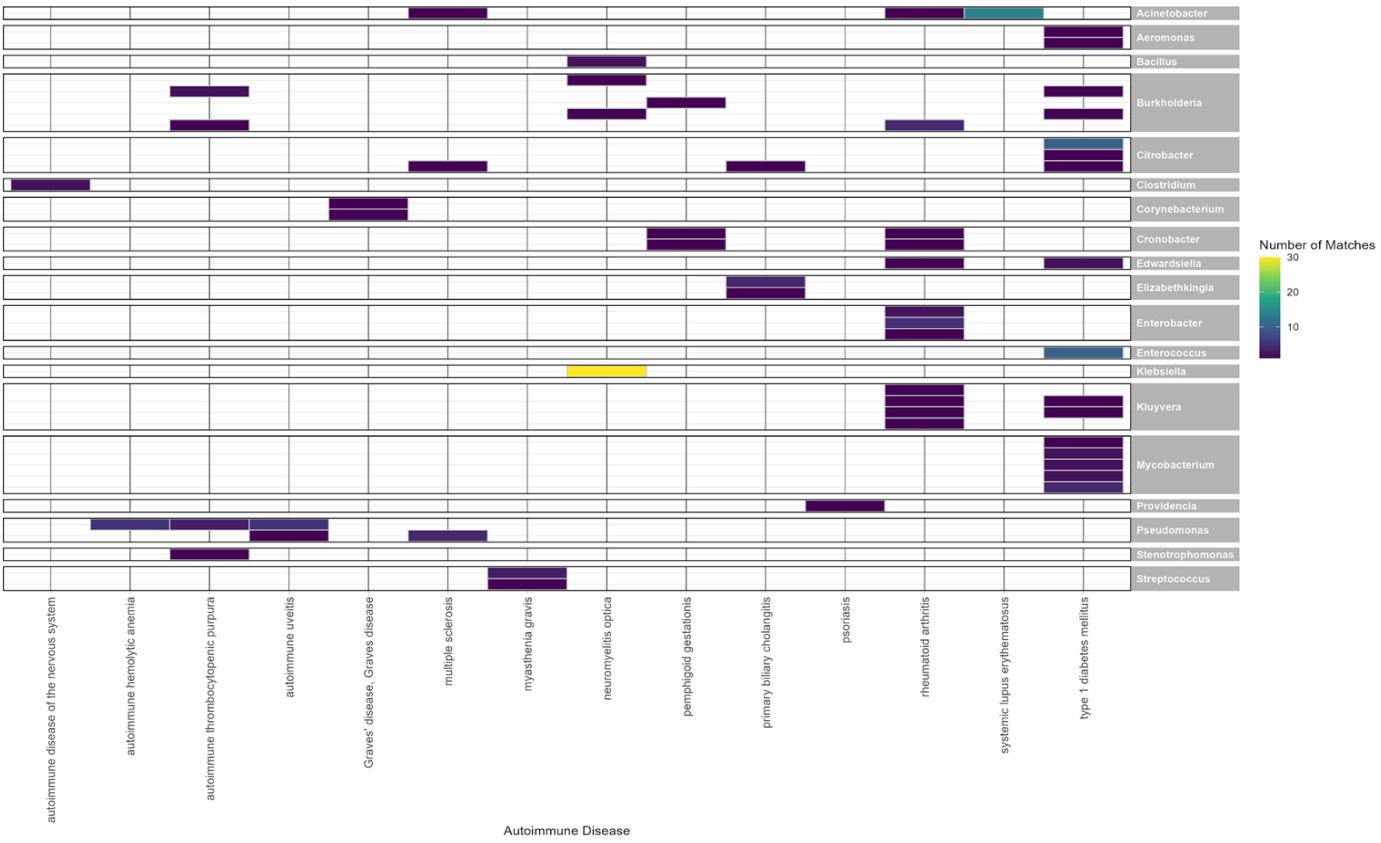


**Figure X. Heatmap showing cross-reactivity potential between autoimmune diseases and pathogen species.** Each tile represents the number of pathogen proteins from a specific species matched to a given autoimmune disease. The plot is faceted by pathogen genus, with one row per species. Tile color indicates the total number of matches.