Contents lists available at ScienceDirect

ImmunoInformatics

journal homepage: www.journals.elsevier.com/immunoinformatics

Epitopedia: identifying molecular mimicry between pathogens and known immune epitopes

Christian A Balbin[a,#,*], Janelle Nunez-Castilla[a,#], Vitalii Stebliankin[b], Prabin Baral[c], Masrur Sobhan[d], Trevor Cickovski[b], Ananda Mohan Mondal[b,d,e], Giri Narasimhan[b,e], Prem Chapagain[c,e], Kalai Mathee[e,f], Jessica Siltberg-Liberles[a,e,*]

[a] *Department of Biological Sciences, College of Arts, Science and Education, Florida International University, Miami, United States*
[b] *Bioinformatics Research Group (BioRG), Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, United States*
[c] *Department of Physics, College of Arts, Science and Education, Florida International University, Miami, United States*
[d] *Machine Learning and Data Analytics Group (MLDAG), Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, United States*
[e] *Biomolecular Sciences Institute, Florida International University, Miami, United States*
[f] *Department of Human and Molecular Genetics, Herbert Wertheim College of Medicine, Florida International University, Miami, United States*

ABSTRACT

Upon infection, foreign antigenic proteins stimulate the host's immune system to produce antibodies targeting the pathogen. These antibodies bind to regions on the antigen called epitopes. Structural similarity (molecular mimicry) of epitopes between an infecting pathogen and host proteins or other pathogenic proteins the host has previously encountered can impact the host immune response to the pathogen and may lead to cross-reactive antibodies. The ability to identify potential regions of molecular mimicry in a pathogen can illuminate immune effects which are especially important to pathogen treatment and vaccine design. Here we present Epitopedia, a software pipeline that facilitates the identification of regions that may exhibit potential three-dimensional molecular mimicry between an antigenic pathogen protein and known immune epitopes as catalogued by the Immune Epitope Database (IEDB). Epitopedia is open-source software released under the MIT license and is freely available on GitHub, including a Docker container with all other software dependencies preinstalled. We performed an analysis describing how various secondary structure states, identity between pentapeptide pairs, and identity between the parent sequences of pentapeptide pairs affects RMSD. We found that pentapeptides pairs in a helical conformation had considerably lower RMSD values than those in extended or coil conformations. We also found that RMSD is significantly increased when pentapeptide pairs are from non-homologous sequences.

1. Introduction

Pathogens present antigenic molecules that can elicit a host immune response. For proteins, an epitope is the portion of the antigen that is recognized and bound by an antibody. Occasionally, pathogen epitopes may share similar chemical and structural properties to unrelated host epitopes, leading to unexpected interactions between the pathogen's epitope and host proteins [1]. Molecular mimicry can also potentially lead to autoimmune disorders where infection with a pathogen can trigger the production of antibodies that mistakenly cross-react with an epitope in a host protein, potentially resulting in a diverse range of autoimmune complications involving both B cell and T cell response [2]. Alternatively, molecular mimicry may lead to heterologous immunity where infection with one pathogen can provide protection against other pathogens that exhibit molecularly similar antigenic proteins [3].

Epitopes can be linear or conformational. Linear epitopes consist of short local sequence stretches while conformational epitopes consist of sequence stretches across the protein sequence that come together in the 3D structure. Prediction of molecular mimicry for conformational epitopes presents a challenge, while the prediction of molecular mimicry at linear epitopes using a sequence-based approach followed by structural comparison is more straightforward. To the best of our knowledge there are currently no computational programs dedicated to the prediction of molecular mimicry of known epitopes using both sequence and 3D structure similarity of peptides, although programs based on sequence similarity searches alone [4] or that map peptides (mimotopes) onto the antigenic protein structure to identify a native epitope [5–8] exist. Further, there are pipelines that identify molecular mimicry in remote homologs on the protein level [9] or that identify molecular mimicry in antibody-binding interfaces [10].

* Corresponding authors.
E-mail addresses: cbalbin@fiu.edu (C.A. Balbin), jliberle@fiu.edu (J. Siltberg-Liberles).

[#] These authors contributed equally to this work.

https://doi.org/10.1016/j.immuno.2023.100023
Received 22 February 2022; Received in revised form 10 January 2023; Accepted 26 January 2023
Available online 27 January 2023

We present Epitopedia, a computational sequence and structure-based pipeline for the prediction of molecular mimicry. Epitopedia identifies structural similarity between short identical sequence fragments in an antigenic protein of interest and experimentally verified linear epitopes found in the Immune Epitope Database (IEDB: www.iedb.org) [11] without considering the homology between the antigenic protein and the source protein of the epitopes. While epitopes vary in length, five amino acid long sequence fragments (pentapeptides) have been described as the minimal immunogenic unit able to form functional interactions with B cells, T cells and MHC molecules [12]. Epitopedia is designed to find molecular mimicry of epitope fragments as short as a pentapeptide but includes longer fragments as possible hits. We include an analysis on the impact from secondary structure when determining structural similarity between pentapeptides to guide the interpretation of molecular mimicry between epitopes and pathogenic proteins. High structural similarity between an epitope and a pathogenic protein is interpreted as potential molecular mimicry and it follows that binding of the same antibody may be possible.

2. Epitopedia implementation

2.1. Internal database generation

Epitopedia utilizes data from the Immune Epitope Database (IEDB) [11], the Protein Data Bank (PDB) [13], and, optionally, the AlphaFold Protein Structure Database [14] for the human proteome [15]. The data are organized into four internal tables (IEDB-FILT, mmcif-seqs, EPI-3D, and 3D-DSSP) stored in a SQLite3 database. IEDB-FILT is derived from a reduced IEDB that only includes the necessary data (epitope sequence, epitope identifier, antigen source sequence, range, accession, organism, etc.) for epitope mimicry search, including the full-length antigen source sequences from all assays available for T Cell, B Cell, and MHC Ligand in IEDB. Based on the epitopes with positive assays from IEDB-FILT, a database for BLASTP (referred to as EPI-SEQ) of linear epitope sequences (mean length of 13 residues) and associated taxonomic origin of the epitopes is generated. Sequences from all PDB structures and human AlphaFold models were extracted and stored in mmcif-seqs. To find structural representatives for the antigen source sequences from IEDB, a sensitive ($s=7.5$) MMseqs2 [16] many-against-many search of antigen source sequences against mmcif-seqs is performed and the results are stored in EPI-3D. For a structural representative to be included in EPI-3D, the MMseqs2 pairwise alignment between the antigen source sequence and the structure sequence must have at least 90% identity to include highly similar homologs and 20% query coverage because many PDB structures are of truncated, partial proteins. Lastly, DSSP [17] is used to determine secondary structure and compute the accessible surface area (ASA) for every residue in each chain in EPI-3D and the results are stored in 3D-DSSP.

2.2. Searching for 1-dimensional molecular mimics

The Epitopedia pipeline is executed with one or more PDB IDs from the same pathogen protein as input. The protein sequence (seqres) is extracted from the input structure and used in a BLASTP search against EPI-SEQ. The BLASTP parameters evalue and max_target_seq are both set to 2,000,000 to avoid discarding hits due to large evalues or reaching the match limit, respectively. The BLAST hits are filtered to only include hits with regions containing 5 or more consecutive, identical amino acids between the query (input protein based on the PDB ID input) and subject (epitope). If a hit meets this requirement in more than one region, the regions are split into subalignments so that one epitope may have >1 region.

Further, to be considered molecular mimics, the regions must have at least 3 consecutive accessible amino acids with a relative accessible surface area (RASA) $>$ 20%, a commonly used cutoff for determining if residues are buried or exposed [18,19]. Based on ASA from 3D-DSSP and the maximum allowed solvent accessibility (MaxASA) values per amino acid as defined in Wilke [20], RASA is calculated according to the equation

$$RASA = ASA/MaxASA$$

Regions meeting these qualifications are considered one dimensional mimics (1D-mimics). Regions that do not meet the aforementioned criteria to be considered a 1D-mimic are discarded.

2.3. Identifying 3-dimensional molecular mimics

For 1D-mimics where the antigen source protein containing the epitope hit is represented in EPI-3D, the structural regions of the input structure corresponding to the 1D-mimic regions are evaluated to ensure that all residues are solved. To avoid missing potential mimics due to regions of missing electron density in an input structure, several structures can simultaneously be used as an input. Further, providing multiple PDB IDs for the same protein as input allows for a conformational ensemble approach to search for structural mimics. The structural fragments of 1D-mimics represented in EPI-3D and the corresponding hit fragment from the input structure are extracted. To complement structural representation of human antigen source proteins in PDB, structural fragments can also be extracted from AlphaFold2 models for the human proteome [15]. Although AlphaFold2 models are used, we refer to them as AlphaFold models from here on after.

TM-align [21] is used to evaluate the structural similarity based on the RMSD for each extracted peptide structure pair based on its BLAST hit pairwise alignment. To ensure that the structural superposition step is in agreement with the peptide pair sequence alignment, the pairwise alignment of the 100% identical 1D-mimic peptide pair is provided to TM-align. Pairs with an RMSD $\leq$ 1Å are considered three dimensional mimics (3D-mimics).

2.4. Handling redundancy and quantifying results

Given the nature of epitopes and IEDB, it is common to have several overlapping epitopes where both the epitope mimic region and the antigen source sequence are identical. Internal accession numbers for all antigen source sequences in IEDB-FILT were assigned to ensure that any two or more identical sequences will have the same internal accession number to allow for filtering of redundancy at the output stage of the pipeline.

Epitopedia outputs results in CSV, JSON, and a simple web interface. The web interface is built using Flask, Bootstrap, and NGL Viewer [22] and provides an interactive visualization of the 3D-mimic region in both the input and epitope-containing antigen source protein. For each run, with N inputs, the distribution of RMSD values for the 3D-mimics is plotted as a histogram, with grey lines denoting the points of -1, 0, 1 standard deviations. The RMSD for each hit is denoted with a red line in the RMSD histogram. The Z-score for the hit is also computed, allowing for a comparative assessment of the hit quality against other hits for a particular run. An additional score termed EpiScore is calculated by dividing the mimic length by the RMSD (length of alignment/RMSD) to emphasize the significance of longer mimics. For example, given several mimics of varying length with the same RMSD, a longer mimic would have a higher EpiScore than a shorter mimic. Further, the EpiScore can reflect a more notable hit for a longer mimic with a higher RMSD than a shorter mimic with a lower RMSD. Thus, a higher EpiScore represents a more remarkable hit. As for RMSD, the EpiScore distribution for each run, shown as a histogram with the hit in red, is included in the web interface.

2.5. User customization

For each provided input structure, the following main steps allow for customization of the run.

For the BLASTP search in *Step 1* (Fig. 1), the user can specify a taxonomy filter for a focused search. With the taxonomy filter, epitopes from the specified taxonomic id will be excluded from the search.

For extracting potential epitope hits based on the input structure in *Step 2*, the minimum span length of an identical hit and the minimum accessibility of the hit in the input structure can be specified, with default values set to 5 and 3 residues, respectively. The user determines the cutoff for RASA, with the default set to 0.2. The sequence motifs from the epitope hits that meet span length and accessibility cutoffs are considered 1D-mimics, because although they are valid epitope hits based on the input structure, the structure of the epitope hit fragment is yet unknown. The structural fragments corresponding to the motif of each 1D-mimic are excised from the input structure.

In *Step 3*, for epitope hits corresponding to 1D-mimics from *Step 2*, the PDB structure of their antigen source protein is extracted from EPI-3D, if such a structure exists. Fragments matching the motifs of the 1D-mimics are excised for later comparison to the corresponding motif of each 1D-mimic from the input structure. Further, accessibility of the residues in the motifs is extracted from 3D-DSSP based on the whole protein structure.

Similarly, the user can choose to extract representative structures from an AlphaFold model of the human proteome [15] based on EPI-3D in *Step 4*. The user can specify the confidence level of the AlphaFold models to consider using a motif (local) and a protein (global) confidence score. Both scores are based on pLDDT, which is the primary confidence score reported for AlphaFold models [23]. For the motif confidence score (m-pLDDT), no residue within the 1D-mimic motif can be below the cutoff. For the protein confidence score (p-pLDDT), the average of pLDDT for the entire model cannot be below the cutoff. The defaults are set to 0.9 and 0.7 for m-pLDDT and p-pLDDT, respectively. Structural fragments matching the motifs of the 1D-mimics are excised for later comparison to the corresponding motif of each 1D-mimic from the input structure. Further, accessibility of the residues in the motifs is extracted from 3D-DSSP based on the whole AlphaFold model.

In *Step 5*, structural comparisons of each motif fragment from the input structure to the corresponding fragments from *Step 3* or *Step 4* are performed using TM-align for the exact pairwise sequence alignment [21]. TM-score and RMSD are reported. However, because only short structural fragments are compared, the TM-score is not meaningful, while the RMSD of the structural alignment and agreement in RASA (based on the whole structural context) are meaningful. The user can set an RMSD cutoff for hits to be reported but the default is no RMSD cutoff.

In *Step 6*, all results for all input structures are compiled into a list. The EpiScore and Z-scores are computed. Hits with RMSD of at most 1Å are considered 3D-mimics. For the 3D-mimics, a web interface output is generated. The web interface includes the settings used to execute Epitopedia and basic information about the motif in the input structure, the epitope it mimics, and the antigen source protein in addition to RMSD, accessibility, EpiScore, Z-scores, and a link to a visualization of the results (Fig. 2). For motifs with a 3D-mimic, the best hit is shown but the other hits are included under a dropdown menu. Structural visualization of 3D-mimics highlights the location of each mimic in the input structure and in the antigen source structural representative (Fig. 3).

3. Epitopedia demonstration

To demonstrate an Epitopedia run, we provide an example using an electron microscopy structure of the SARS-CoV-2 Spike protein (PDB ID: 6VXX, chain A [24]) as input (Fig. 4). The taxid-filter flag with a taxid of 11118 was utilized to ensure neither the input protein nor other

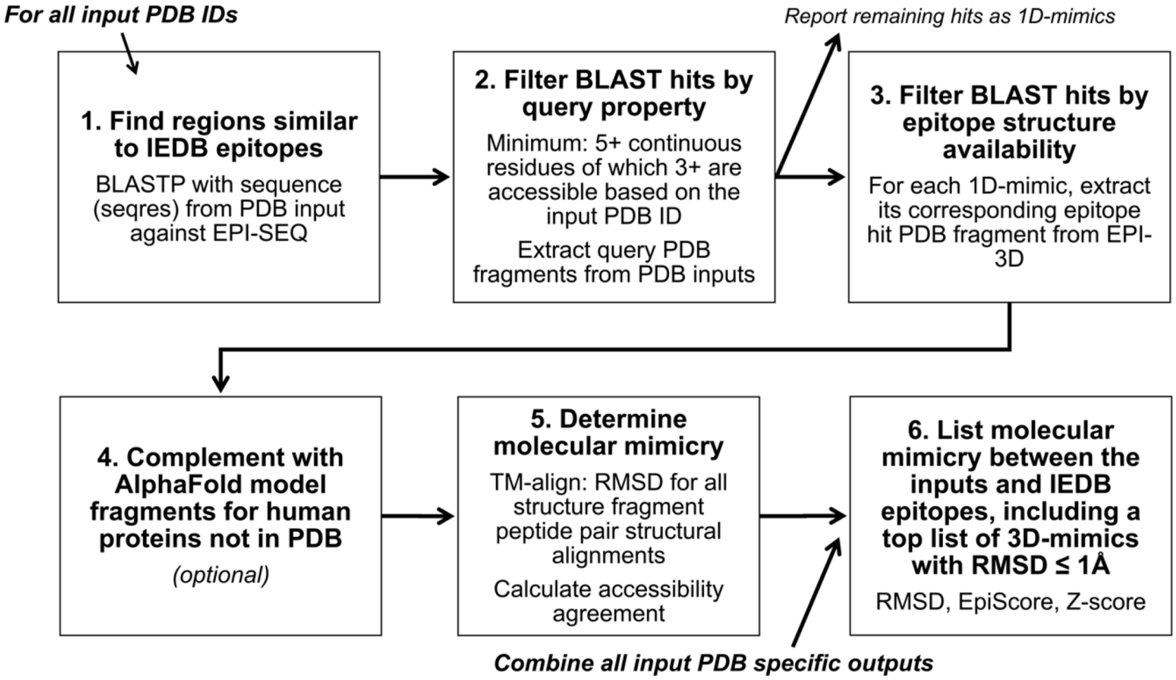
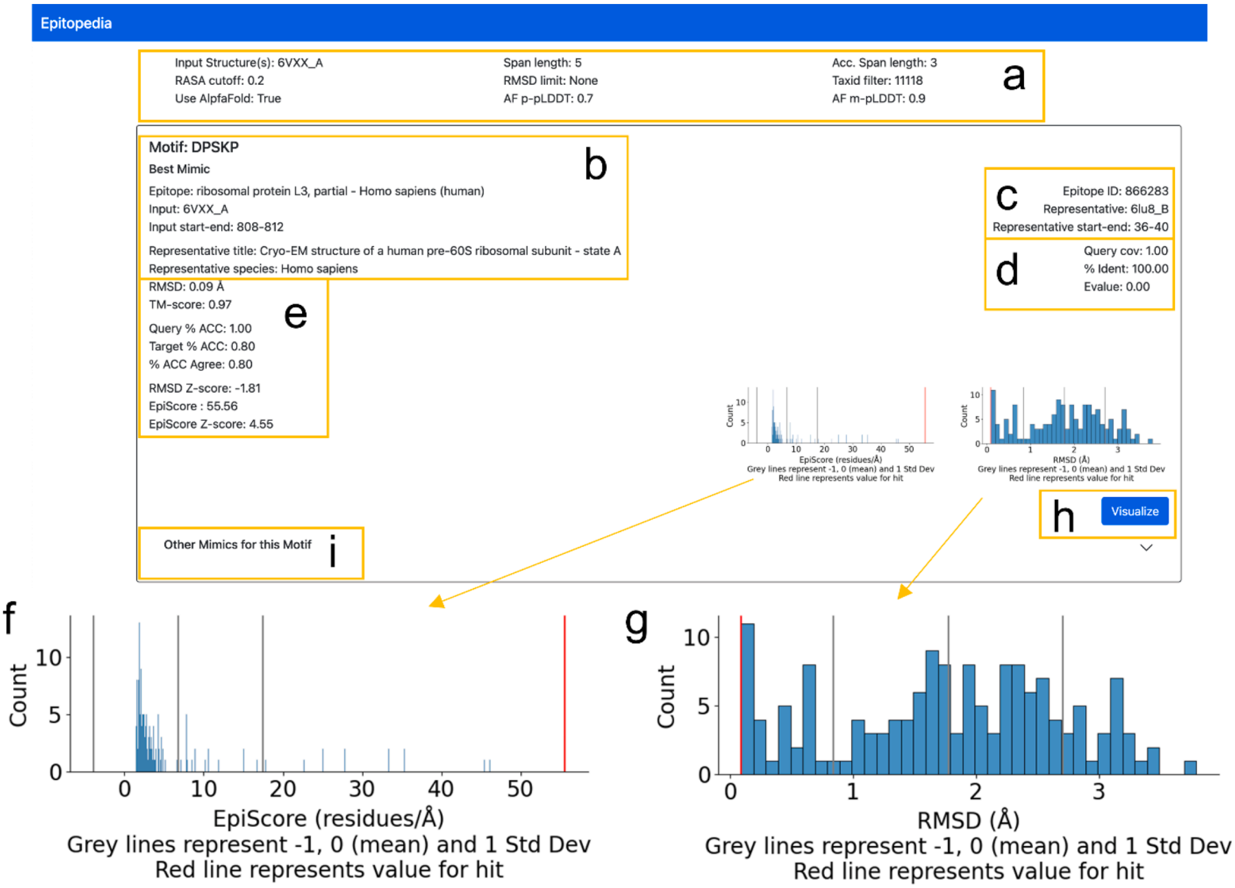

For all input PDB IDs
Report remaining hits as 1D-mimics
1. Find regions similar to IEDB epitopes
BLASTP with sequence (seqres) from PDB input against EPI-SEQ
2. Filter BLAST hits by query property
Minimum: 5+ continuous residues of which 3+ are accessible based on the input PDB ID
Extract query PDB fragments from PDB inputs
3. Filter BLAST hits by epitope structure availability
For each 1D-mimic, extract its corresponding epitope hit PDB fragment from EPI-3D
4. Complement with AlphaFold model fragments for human proteins not in PDB
(optional)
5. Determine molecular mimicry
TM-align: RMSD for all structure fragment peptide pair structural alignments
Calculate accessibility agreement
6. List molecular mimicry between the inputs and IEDB epitopes, including a top list of 3D-mimics with RMSD ≤ 1Å
RMSD, EpiScore, Z-score
Combine all input PDB specific outputs

Fig. 1. Overview of Epitopedia. Epitopedia is initiated with one or more PDB structures from the same protein as input. In *Step 1*, a BLASTP search against linear epitope sequences in EPI-SEQ is performed with the corresponding sequence (seqres) from each PDB input as query. In *Step 2*, BLASTP hits that include sequence fragments from the query that do not contain at least 5 consecutive identical amino acids and where less than 3 amino acids are surface accessible based on the input structure are discarded. For the remaining hits, the PDB fragment is extracted from the input structure. These are considered 1D-mimics. In *Step 3*, structural fragments from the hits from EPI-SEQ that correspond to the 1D-mimics are extracted from PDB structural representatives of the source antigens. In *Step 4* (optional), for hits against epitopes in human source antigens that are not represented in PDB, structural fragments are extracted from AlphaFold models for regions with a certain confidence level (specified by the user). In *Step 5*, TM-align is used to calculate the RMSD of the structural alignment of the BLAST hit fragment or peptide pairs. In *Step 6*, RMSD results for all fragment pairs for all inputs for the run are combined. EpiScore (length of alignment/RMSD) and RMSD histograms are generated, and Z-scores are calculated based on the whole run. A top list of fragment pairs with RMSD $\leq$ 1Å is created. These fragment pairs are referred to as 3D-mimics.


Epitopedia

Input Structure(s): 6VXX_A
RASA cutoff: 0.2
Use AlpfaFold: True
Span length: 5
RMSD limit: None
AF p-pLDDT: 0.7
Acc. Span length: 3
Taxid filter: 11118
AF m-pLDDT: 0.9

Motif: DPSKP
Best Mimic
Epitope: ribosomal protein L3, partial - Homo sapiens (human)
Input: 6VXX_A
Input start-end: 808-812
Representative title: Cryo-EM structure of a human pre-60S ribosomal subunit - state A
Representative species: Homo sapiens

Epitope ID: 866283
Representative: 6lu8_B
Representative start-end: 36-40

Query cov: 1.00
% Ident: 100.00
Evalue: 0.00

RMSD: 0.09 Å
TM-score: 0.97
Query % ACC: 1.00
Target % ACC: 0.80
% ACC Agree: 0.80
RMSD Z-score: -1.81
EpiScore : 55.56
EpiScore Z-score: 4.55

Other Mimics for this Motif

a
b
c
d
e
h
Visualize
i
f
Count
10
5
0
0
10
20
30
40
50
EpiScore (residues/Å)
Grey lines represent -1, 0 (mean) and 1 Std Dev
Red line represents value for hit
g
Count
10
5
0
0
1
2
3
RMSD (Å)
Grey lines represent -1, 0 (mean) and 1 Std Dev
Red line represents value for hit

Fig. 2. Overview of the Epitopedia web interface for 3D-mimics. For each run, (a) information about the run; (b) the mimic and protein in which the mimic was identified; (c) the epitope and its structural representative; (d) identification of the structural representative with MMseqs2; (e) structural comparison of the mimics including EpiScore, EpiScore Z-Score, and RMSD Z-Score; (f) EpiScore distribution for all structurally represented mimics (blue) during the given run including the EpiScore Z-score (grey), with the current mimic in red; (g) RMSD distribution for all structurally represented mimics (blue) during the given run including the RMSD Z-score (grey), with the location of the current mimic in red; (h) link to 3D visualization of the mimic; (i) and while the Best Mimic is shown from the start, additional mimics for the same motif from the same or different proteins but with higher RMSD are included in a dropdown menu.

Coronavirus proteins were included as mimics (since these are homologous proteins and not mimics). The search for mimic representatives was performed against both PDB and the Human AlphaFold Protein Structure Database with default settings.

The run resulted in 755 1D-mimics, where 297 1D-mimics are structurally represented, of which 93 are only represented in the Human AlphaFold Protein Structure Database. After ensuring that only the best mimic per source sequence progresses, there were 153 mimics, with 66 of them mimicked with an AlphaFold structure. Finally, after filtering the results so that only 3D-mimics with an RMSD $\leq$ 1Å remain and removing redundant hits, there were 27 mimics, of which 11 are mimicked with an AlphaFold structure. Of the 16 3D-mimics from PDB, 13 are from human (such as integrin beta-1), and one each are from *Mycobacterium tuberculosis, Bacillus anthracis*, and Timothy grass (Table S1, Figs. S1-S6). The remaining 11 3D-mimics are from the Human AlphaFold Protein Structure Database [14,15] and thus are all from human epitopes (Table S2). The mimic with the lowest RMSD (0.09 Å) is shown in Figs. 2 and 3. This hit is for a subunit in the human pre-60S ribosome, which is an intracellular protein. Intracellular proteins have generally lower risk of potential cross-reactive autoimmunity. Epitopedia does not differentiate between protein function or cellular localization so each user must critically analyze their results for biological relevance.

We also applied Epitopedia to a different structure of the SARS-CoV-2 Spike protein (PDB ID: 6XR8, chain X [25]) and identified molecular mimics with implications for understanding COVID-19 pathogenesis [26]. Multiple interesting 3D-mimics were found, but in particular, the TQLPP motif from an epitope in human thrombopoietin (hTPO), a secretory protein involved in the regulation of platelet production [27], piqued our interest. Thrombocytopenia, a condition characterized by low platelet levels, has been well-documented in individuals infected with [28] and vaccinated against [29] SARS-CoV-2 [30,31], but the mechanism for developing thrombocytopenia in the SARS-CoV-2 context is unknown. In SARS-CoV-2 Spike, the TQLPP epitope identified by Epitopedia is located on the surface of the N-terminal domain. In hTPO, TQLPP is found at the interface with a neutralizing antibody (PDB ID: 1V7M [32]). Although Spike and hTPO are unrelated proteins, TQLPP is found in a highly similar coil conformation in both (RMSD = 0.61 Å). In silico docking of the antibody from the hTPO structure against multiple Spike conformations demonstrate that the antibody binds comparably to both hTPO and Spike, suggesting the possibility of Spike eliciting the production of cross-reactive antibodies against hTPO.

Altogether, these two examples of Epitopedia results highlight the importance of the input structure for the search. To avoid missing interesting molecular mimics, it is important to use input structures that cover as much of the pathogen protein as possible. In the case of the TQLPP motif, it is located in positions 22-26 in the Spike protein from SARS-CoV-2 and while PDB id: 6XR8 starts at position 14, PDB id: 6VXX starts at position 27 which is why the search with 6VXX did not identify the potential mimicry between Spike and hTPO. Further, for proteins with multiple conformations represented in the PDB, including more than one structure can generate additional hits.

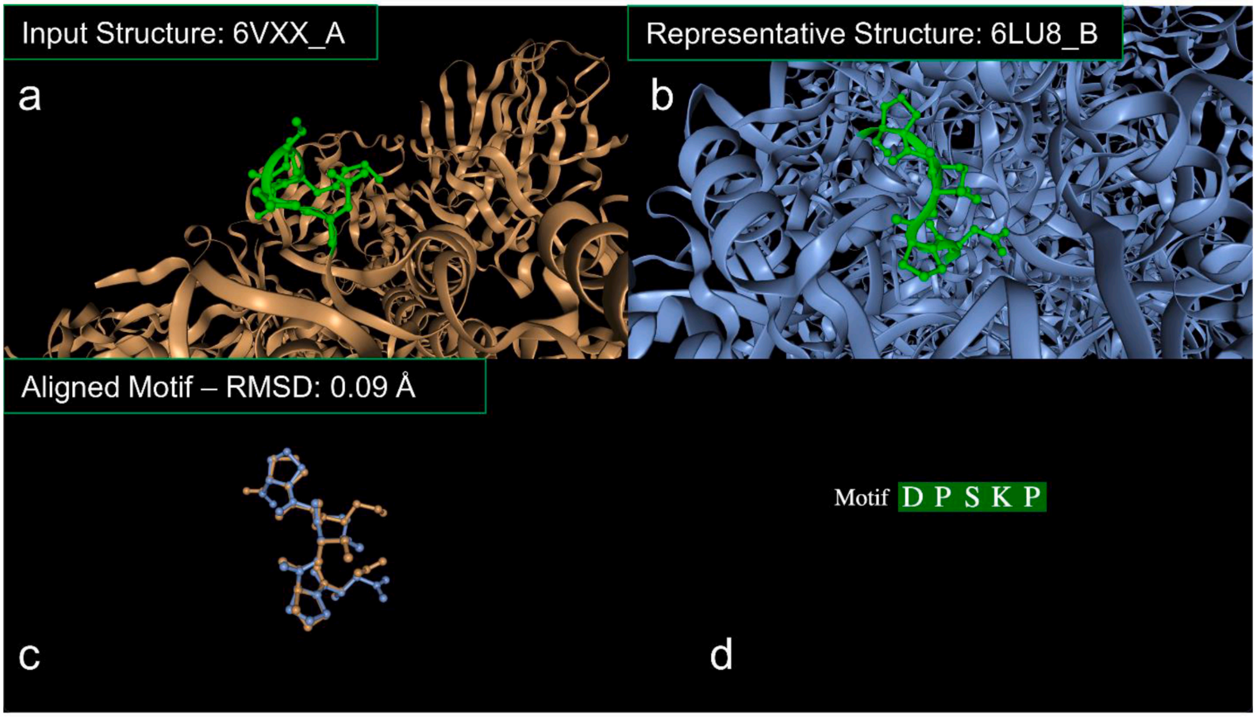

Input Structure: 6VXX_A
a
Representative Structure: 6LU8_B
b
Aligned Motif – RMSD: 0.09 Å
c
Motif D P S K P
d

Fig. 3. Visualization of the mimic pair in 3D. (a) The motif (green) shown in input protein (brown) (b) and in the structural representative protein (blue). (c) The TM-align structural superimposition for the motif in the input protein (brown) and the structural representative (blue). Panels a-c are interactive. (d) The mimic motif is interactive, hovering over a residue in the motif will highlight it in panels a-c.

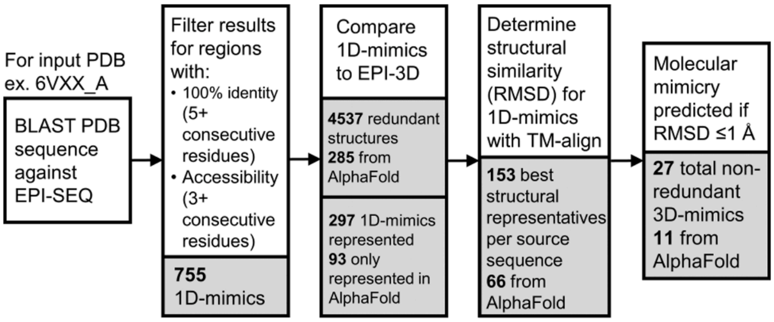

For input PDB ex. 6VXX_A
BLAST PDB sequence against EPI-SEQ
Filter results for regions with:
• 100% identity (5+ consecutive residues)
• Accessibility (3+ consecutive residues)
755 1D-mimics
Compare 1D-mimics to EPI-3D
4537 redundant structures
285 from AlphaFold
297 1D-mimics represented
93 only represented in AlphaFold
Determine structural similarity (RMSD) for 1D-mimics with TM-align
153 best structural representatives per source sequence
66 from AlphaFold
Molecular mimicry predicted if RMSD ≤1 Å
27 total non-redundant 3D-mimics
11 from AlphaFold

Fig. 4. Epitopedia output overview using PDB 6VXX, chain A as input. For detailed output see example_output folder on the GitHub repository.

4. Pentapeptide structural space analysis

To provide guidance on how to interpret structural mimicry based on RMSD for the 3D-mimic pentapeptide fragment pairs identified by Epitopedia, we performed an investigation of RMSD for random pentapeptide pairs for the three main secondary structure states helix, extended, and coil from any sequence pair regardless of sequence similarity and for sequence pairs with low sequence similarity representing non-homologous proteins.

4.1. Methods

To understand how secondary structure state and sequence identity affect the distribution of RMSD values for pentapeptide pairs, an analysis of RMSD distributions of pentapeptide pairs across various secondary structure states and pentapeptide sequence identity levels was performed.

All possible pentapeptides based on PDB structures were generated and annotated with a DSSP secondary structure state reduction based on 3D-DSSP. The DSSP state reduction was performed such that if all residues in a pentapeptide were classified as turn (T), bend (S) or none (-), the pentapeptide was labeled coil, if all residues were strand (E) or beta-bridge (B) the pentapeptide was labeled extended, and if all residues were alpha helix (H), 3-10 helix (G), or pi-helix (I) the pentapeptide was labeled helix. Any pentapeptides that did not fit into one of these 3 categories were discarded.

Around 1,000 pentapeptide pairs (Table S3) were generated for each secondary structure state per identity level (0%, 20%, 40%, 60%, 80%, and 100%) from the labeled pentapeptide database described above. The number of pentapeptide pairs per category is not exactly the same across all categories because matches of a pentapeptide against itself (same PDB ID) are discarded. The pentapeptide regions were extracted from the parent structures using GEMMI [33] and superposed using TM-align [21], with a fixed alignment as described for the Epitopedia implementation above.

To reduce the influence that parent sequence homology may have on the above analysis, we performed a similar analysis starting with 2,000 pentapeptides for each secondary structure state per identity level. Here, an added filtering step was performed to ensure that the parent sequences of the pentapeptide pairs were no more than 30% identical according to a local pairwise Smith-Waterman alignment of the parent sequences generated with EMBOSS Water [34]. The 30% identity cutoff between the parent sequences was set to include only unrelated or at most remotely related parent sequence pairs. Pentapeptide matches where the identity filter could not be enforced were discarded, thus, the number of pentapeptide pairs per category is not exactly the same across categories. For instance, if a query pentapeptide had been paired with over 100 other pentapeptides to generate a pentapeptide pair, yet a pentapeptide pair with a parent sequence identity of less than 30% was not found, the query pentapeptide was discarded. This scenario disproportionately affected pentapeptide pairs with higher pentapeptide identity, as there is a lower chance of parent sequences having less than 30% identity as the pentapeptide pair identity increases. In total, all pentapeptide identity and secondary structure combinations have greater than 900 pentapeptide pairs (Table S3).

Statistical comparisons were performed with Mann Whitney U using *SciPy* [35]. Alpha values were corrected for multiple comparisons using simple Bonferroni correction. For a confidence level of 99%:

$$corrected\ alpha = \frac{0.01}{N\ pairwise\ comparisons}$$

4.2. Results

An analysis was performed to better understand how the RMSD distribution for pentapeptides pairs varies with differing pentapeptide pair sequence identity, parent sequence identity and secondary structure state. For the first analysis that did not consider the percent identity of the parent sequences for a pentapeptide pair, a decrease in the median RMSD is observed at the 100% identity levels (Fig. 5, Table 1).

For helix pentapeptide pairs, the median RMSD for the 0% to 80% pentapeptide identity levels is 0.20-0.22Å, while at the 100% identity level the median is 0.13Å, which is a statistically significant decrease when compared to all other identity levels for the helical state (Table 2). For extended pentapeptide pairs, the median RMSD for the 0% to 80% pentapeptide identity levels is 0.69-0.84Å, while at the 100% identity level the median is 0.14Å. This is a statistically significant decrease when compared to all other identity levels for the extended state (Table 2). Lastly, for coil pentapeptide pairs, the median RMSD for the 0% to 80% pentapeptide identity levels is 1.79-1.95Å, while at the 100% identity level the median is 0.31Å. This large decrease of ~1.5Å is statistically significant when compared to all other identity levels for coil (Table 2).

Table 1
Median RMSD values resulting from RMSD distribution for structural space analysis of pentapeptide pairs of various identity levels and secondary structure categories shown in Fig. 5.

Pentapeptide % Identity	Median RMSD (Å)					
	No Parent Identity Filter			30% Parent Identity Filter		
	Helix	Extended	Coil	Helix	Extended	Coil
0	0.22	0.84	1.95	0.22	0.83	1.91
20	0.21	0.82	1.88	0.21	0.80	1.90
40	0.21	0.75	1.87	0.21	0.77	1.88
60	0.20	0.74	1.85	0.21	0.74	1.85
80	0.21	0.69	1.79	0.21	0.76	1.79
100	0.13	0.14	0.31	0.20	0.63	1.74

For the follow-up analysis we enforced a maximum 30% parent sequence identity filter to better resemble molecular mimics from unrelated protein pairs. To ensure that the pentapeptide pairs were from proteins that are not closely related, we performed local alignments and extracted the percent sequence identity and query cover for each parent sequence pair. By design, no parent sequence pair has a pairwise sequence identity above 30%, with a median around 20% (Fig. 5). The query cover for the parent sequence pair alignments is low, with a median of around 20% (Fig. 5). For these pairwise sequence alignments with 20% sequence identity and query cover, we can assume that these are primarily non-homologous parent sequence pairs although some remote homologs may be included in this dataset.

For the pentapeptide pairs from these non-homologous sequence pairs, the sharp decrease in the median RMSD at the 100% pentapeptide identity level has faded for extended and coil conformations (Fig. 5). For helix pentapeptide pairs, the median RMSD at the 0% to 100% pentapeptide identity level is 0.20-0.22Å. Only the 100% vs 0% identity level comparison yields a statistically significant difference for the helix pentapeptide pairs (Table 2). For extended pentapeptide pairs, the median RMSD at the 0% to 100% pentapeptide identity level is 0.63-0.83Å. For coil pentapeptide pairs, the median RMSD at the 0% to 100% pentapeptide identity level is 1.74-1.91Å. For extended and coil pentapeptide pairs, the 100% identity level is significantly different when compared against every other identity level except for one comparison, 80% vs 100% in the coil state (Table 2).

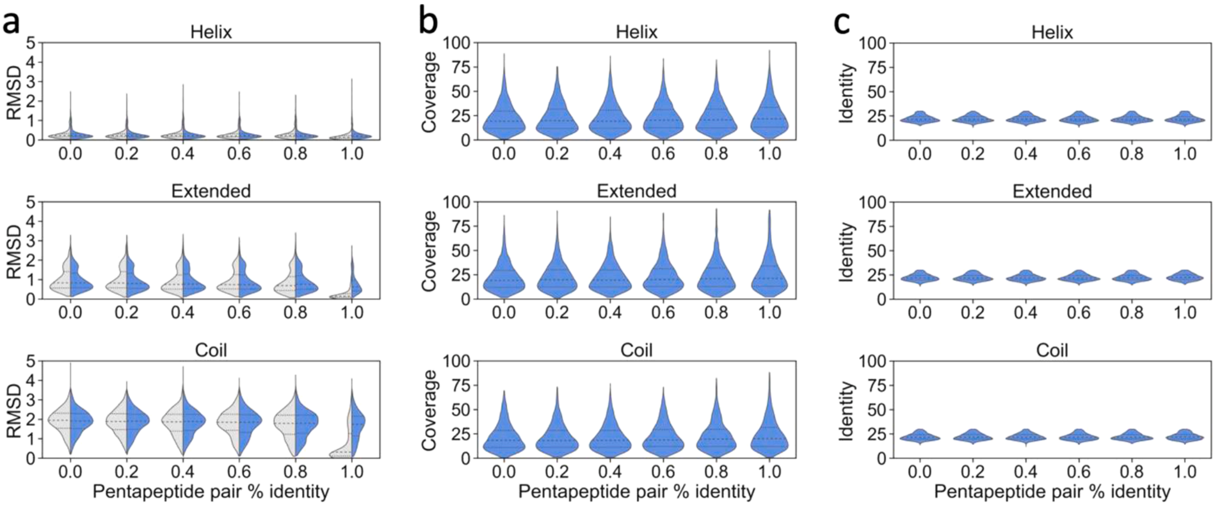


Fig. 5. (a) Violin plots of the resulting RMSD distribution from pentapeptide structure analysis. The distributions for the analysis without the 30% parent sequence identity filter are shown in grey while the corresponding distributions for the pentapeptides from the 30% parent sequence identity set are shown in blue. (b) Violin plots showing the distribution of query coverage between the parent sequences for pentapeptide pairs at various identity levels and secondary structure categories. (c) Violin plots showing the distribution of pairwise identity between the parent sequences for pentapeptide pairs at various identity levels and secondary structure categories.

Table 2
Comparisons across pentapeptide identity levels for the same secondary structure category for the no parent identity identity filter (may include closely related protein pairs) and 30% parent filter datasets (unrelated or at most remotely related protein pairs), respectively. Comparisons with statistically significant differences are denoted *.[1]

Pentapeptide % Identities Compared		No Parent Identity Filter *p-value* Helix	Extended	Coil	30% Parent Identity Filter *p-value* Helix	Extended	Coil
0	20	0.197694	0.578304	0.070875	0.084507	0.107303	0.464566
0	40	0.377122	0.000931	0.001582	0.212098	0.002897	0.013426
0	60	0.076141	0.000368	0.000528	0.15162	1.64E-06*	7.41E-06*
0	80	0.313838	3.52E-12*	5.1E-09*	0.0284	9.37E-08*	1.97E-10*
0	100	8.4E-59*	1.3E-212*	5.5E-156*	5.63E-06*	2.77E-31*	3.33E-17*
20	40	0.688416	0.006825	0.145049	0.636335	0.170195	0.083497
20	60	0.634811	0.00257	0.092602	0.779538	0.001278	0.000172
20	80	0.791704	1.43E-10*	2.84E-05*	0.639867	0.000168	1.84E-08*
20	100	3.63E-52*	2.8E-210*	2.3E-149*	0.002963	4.31E-25*	6.02E-15*
40	60	0.362381	0.785088	0.797398	0.838454	0.06416	0.038889
40	80	0.891271	7.14E-05*	0.006063	0.35334	0.016001	8.83E-05*
40	100	1.27E-54*	4.5E-203*	1.9E-141*	0.00072	2.59E-20*	2.54E-10*
60	80	0.457191	0.000238	0.011388	0.457478	0.571466	0.067424
60	100	1.06E-50*	7.3E-200*	2.5E-140*	0.001357	1.51E-14*	6.59E-06*
80	100	3.08E-53*	2.2E-174*	1.5E-124*	0.01095	5.00E-13*	0.003258

[1] Based on simplified Bonferroni correction at 99% confidence level, corrected alpha = 0.000111.

When comparing the same identity level for the pentapeptide pairs across the set with no parent sequence identity filter and the set with the 30% sequence identity filter, we found that the pairwise parent sequence identity has an impact on the RMSD for identical pentapeptide pairs in the helical state, but not for the less identical peptide pairs (Table 3). This pattern is shared for the coil state, but for the extended state, the pairwise parent sequence identity seems to impact RMSD for identical and 80% identical pentapeptides (Table 3).

Altogether, this analysis shows that for pentapeptides, the secondary structure state is important to consider when identifying molecular mimics using RMSD for random proteins. We used TM-align to calculate RMSD and this method, like many others, calculates RMSD based on the spatial coordinates for C-alpha in each amino acid residue. Our observation that pentapeptide pairs in a helical state have lower RMSD is not surprising given the regular geometry of the α-helix. For identical pentapeptide pairs in extended and coil conformations, the median RMSD for the non-homologous parent sequences are 0.63Å and 1.74Å, respectively, compared to 0.20Å for helix (Table 1).

4.3. Guidelines

Our interpretation, as far as molecular mimicry goes, is that mimics with identical sequences in α-helices are likely to appear very similar if they are oriented the same way in their parent proteins. As such, they are likely to be able to participate in similar interactions with, for example, an antibody. Mimics with identical sequences with low RMSDs, approaching the median RMSD of the unfiltered set (Table 1), are likely to present a similar interaction interface, if oriented similarly. A pentapeptide in a helix, given its winding structure, is relatively small while a pentapeptide in the extended or coil conformation may present a larger accessible area.

Table 3
Comparisons between pentapeptide identity levels for the same secondary structure category for the 30% parent identity filter (likely unrelated protein pairs) vs the no parent identity filter (may be related) dataset. Comparisons with statistically significant differences are denoted *.[1]

Pentapeptide % Identities Compared Filter	No Filter	30% Parent Identity Filter vs no Parent Identity Filter *p-value* Helix	Extended	Coil
0	0	0.438284984	0.56859737	0.1896511
20	20	0.381648164	0.255207316	0.7826223
40	40	0.445432037	0.339928559	0.5969043
60	60	0.091388265	0.844626827	0.4058797
80	80	0.856800634	0.000269319*	0.5128634
100	100	3.41E-55*	4.83E-171*	1.93E-131*

[1] Based on simplified Bonferroni correction at 99% confidence level, corrected alpha = 5.56E-04.

Pathogen proteins that mimic known epitopes in antigenic proteins may stimulate the production of cross-reactive antibodies that can interact with the pathogen protein as well as the human antigen. Pathogen proteins that mimic known epitopes in other pathogens may trigger an immune memory that could lead to protective immunity or complex immune effects such as anti-body dependent enhancement. While Epitopedia does not differentiate between epitope types (B cell, T cell, or MHC) in IEDB-FILT, the user is encouraged to parse the output for the epitope types of interest.

Conclusion

Here, we have developed Epitopedia, a pipeline for the discovery of potential molecular mimics of immune epitopes found in IEDB. Importantly, Epitopedia is designed to only predict molecular mimicry for linear epitopes, that are continuous in sequence, as opposed to conformational epitopes, that are discontinuous in sequence and come together in three-dimensional space. As such, molecular mimics found in conformational epitopes cannot be identified using our approach. Additionally, Epitopedia is reliant on publicly available data found in both IEDB and PDB and cannot predict instances of molecular mimicry *de novo*. Molecular mimics that are not yet found in IEDB or PDB will not be identified by Epitopedia. Furthermore, relying on public databases can lead to biased results because proteins with greater perceived relevance (e.g. those involved in more common human diseases) are more likely to be well-studied and thus have functional and structural information deposited in these databases, while other proteins remain underrepresented. PDB is also biased towards proteins that lack intrinsic disorder and the more stable conformation of a dynamic protein. Therefore, Epitopedia may not predict molecular mimics in conformationally flexible regions. Importantly, results produced by Epitopedia are only predictions, subject to both false positives and negatives. It is critical to further investigate this output with both literature searches and experimental validation.

Epitopedia can facilitate our understanding of how pathogens may interfere with the known epitopes from the human proteome and also known epitopes from other species. Epitopes shared between pathogens can impact immune responses for secondary infections and identification of mimics of epitopes can provide insights to the mechanism behind the widely differing clinical manifestations and complications of infection with certain pathogens, such as SARS-CoV-2. Identification of molecular mimicry between known epitopes from the human proteome and a human pathogen protein can provide clues to the autoimmune potential of an infection caused by the pathogen. Further, by pinpointing regions in the pathogen's proteome that may cause an autoimmune response if a cross-reactive antibody is created against it, these regions can be avoided in future vaccine design. Lastly, by highlighting which human proteins may be at risk for autoimmune targeting in response to a pathogen infection, therapeutics to counteract autoimmune effects can be used (or developed). Epitopedia provides a starting point for generating a better understanding of the autoimmune potential of pathogens and can benefit large-scale data mining and experimental in-vitro and in-vivo design to solve autoimmune conundrums in infectious disease.

Funding

This work was partially supported by the National Science Foundation under Grant No. 2037374.

Author contributions

J.S.-L., G.N., P.C., K.M., T.C., A.M.M. conceived the overall approach. J.S.-L., J.N.-C. conceptualized the Epitopedia pipeline. J.N.-C., C.A.B., J.S.-L. designed and developed the method and approach. J.N.-C. created a pilot of the core pipeline. C.A.B. wrote the code to implement the pipeline. J.N.-C., C.A.B., J.S.-L. analyzed data and performed visualization. J.S.-L supervised the project. All authors discussed the pipeline and provided feedback. J.N.-C., C.A.B., J.S.-L. wrote the manuscript. All authors read and commented the manuscript.

Data and software availability

Epitopedia is primarily written in Python and relies on established software and databases. Epitopedia is available at https://github.com/cbalbin-bio/Epitopedia under the opensource MIT license and also as a docker container at https://hub.docker.com/r/cbalbin/epitopedia.

Supplementary Materials

Supplementary tables S1-S3 and figures S1-S6

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Link under Data and software availability at the end of the manuscript.

Acknowledgments

We thank Teresa Liberatore for discussions.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.immuno.2023.100023.

References

[1] Getts DR, Chastain EM, Terry RL, Miller SD. Virus infection, antiviral immunity, and autoimmunity. Immunol Rev 2013;255(1):197–209. https://doi.org/10.1111/IMR.12091.

[2] Cusick MF, Libbey JE, Fujinami RS. Molecular mimicry as a mechanism of autoimmune disease. Clin Rev Allergy Immunol 2012;42(1):102–11. https://doi.org/10.1007/s12016-011-8294-7.

[3] Agrawal B. Heterologous immunity: role in natural and vaccine-induced resistance to infections. Front Immunol 2019:2631. https://doi.org/10.3389/FIMMU.2019.02631. 0.

[4] Ras-Carmona A, Lehmann AA, Lehmann PV, Reche PA. Prediction of B cell epitopes in proteins using a novel sequence similarity-based method. Sci Rep 2022;12(1):1–9. https://doi.org/10.1038/s41598-022-18021-1. *2022 12:1*.

[5] Chen W, Guo WW, Huang Y, Ma Z. Pepmapper: A collaborative web tool for mapping epitopes from affinity-selected peptides. PLoS One 2012;7(5):37869. https://doi.org/10.1371/journal.pone.0037869.

[6] Huang YX, Bao YL, Guo SY, Wang Y, Zhou CG, Li YX. Pep-3D-Search: A method for B-cell epitope prediction based on mimotope analysis. BMC Bioinf 2008;9. https://doi.org/10.1186/1471-2105-9-538.

[7] Mayrose I, Penn O, Erez E, Rubinstein ND, Shlomi T, Freund NT, Bublil EM, Ruppin E, Sharan R, Gershoni JM, Martz E, Pupko T. Pepitope: Epitope mapping from affinity-selected peptides. Bioinformatics 2007;23(23):3244–6. https://doi.org/10.1093/bioinformatics/btm493.

[8] Negi SS, Braun W. Automated detection of conformational epitopes using phage display peptide sequences. Bioinf Biol Insights 2009;2009(3):71–81. https://doi.org/10.4137/bbi.s2745.

[9] Armijos-Jaramillo V, Espinosa N, Vizcaıno K, Santander-Gordon D. A novel in silico method for molecular mimicry detection finds a formin with the potential to manipulate the maize cell cytoskeleton. Mol Plant-Microbe Interact : MPMI 2021;(7):34. https://doi.org/10.1094/MPMI-11-20-0332-R.

[10] Stebliankin V, Baral P, Balbin C, Nunez-Castilla J, Sobhan M, Cickovski T, Mohan Mondal A, Siltberg-Liberles J, Chapagain P, Mathee K, Narasimhan G. EMoMiS: a pipeline for epitope-based molecular mimicry search in protein structures with applications to SARS-CoV-2. Biorxiv 2022. https://doi.org/10.1101/2022.02.05.479274.

[11] Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B. The immune epitope database (IEDB): 2018 update. Nucleic Acids Res 2019;47(D1):D339–43. https://doi.org/10.1093/nar/gky1006.

[12] Kanduc D. Pentapeptides as minimal functional units in cell biology and immunology. Curr Protein Pept Sci 2013;14(2):111–20. https://doi.org/10.2174/1389203711314020003.

[13] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235–42. https://doi.org/10.1093/nar/28.1.235.

[14] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Velankar S. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 2022;50(D1):D439–44. https://doi.org/10.1093/NAR/GKAB1061.

[15] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A, Velankar S, Kleywegt GJ, Bateman A, Evans R, Pritzel A, Figurnov M, Ronneberger O, Bates R, Kohl SAA, Hassabis D. Highly accurate protein structure prediction for the human proteome. Nature 2021;2021:1–9. https://doi.org/10.1038/s41586-021-03828-1.

[16] Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 2017;35(11):1026–8. https://doi.org/10.1038/nbt.3988. *2017 35:11*.

[17] Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–637. https://doi.org/10.1002/bip.360221211.

[18] Chen H, Zhou HX. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res 2005;33(10):3193. https://doi.org/10.1093/NAR/GKI633.

[19] Savojardo C, Manfredi M, Martelli PL, Casadio R. Solvent accessibility of residues undergoing pathogenic variations in humans: from protein structures to protein sequences. Front Mol Biosci 2021;7:460. https://doi.org/10.3389/FMOLB.2020.626363/BIBTEX.

[20] Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilites of residues in proteins. PLoS One 2013;8(11). https://doi.org/10.1371/journal.pone.0080635.

[21] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33(7):2302–9. https://doi.org/10.1093/NAR/GKI524.

[22] Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlic A, Rose PW. NGL viewer: web-based molecular graphics for large complexes. Bioinformatics 2018;34(21):3755–8. https://doi.org/10.1093/BIOINFORMATICS/BTY419.

[23] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature 2021;2021:1–11. https://doi.org/10.1038/s41586-021-03819-2.

[24] Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 2020;181(2):281–92. https://doi.org/10.1016/J.CELL.2020.02.058. e6.

[25] Cai Y, Zhang J, Xiao T, Peng H, Sterling SM, Walsh RM, Rawson S, Rits-Volloch S, Chen B. Distinct conformational states of SARS-CoV-2 spike protein. Science 2020;(6511):369. https://doi.org/10.1126/science.abd4251.

[26] Nunez-Castilla J, Stebliankin V, Baral P, Balbin CA, Sobhan M, Cickovski T, Mondal AM, Narasimhan G, Chapagain P, Mathee K, Siltberg-Liberles J. Potential autoimmunity resulting from molecular mimicry between SARS-CoV-2 spike and human proteins. Viruses 2022;14(7):1415. https://doi.org/10.3390/V14071415.

[27] Varghese LN, Defour J-P, Pecquet C, Constantinescu SN. The thrombopoietin receptor: structural basis of traffic and activation by ligand, mutations, agonists, and mutated calreticulin. Front Endocrinol 2017;8(MAR):1. https://doi.org/10.3389/FENDO.2017.00059.

[28] Yang X, Yang Q, Wang Y, Wu Y, Xu J, Yu Y, Shang Y, Lillicrap D. Thrombocytopenia and its association with mortality in patients with COVID-19. J Thromb Haemost 2020;18:1469–72. https://doi.org/10.1111/jth.14848.

[29] Helms JM, Ansteatt KT, Roberts JC, Kamatam S, Foong KS, Labayog JMS, Tarantino MD. Severe, refractory immune thrombocytopenia occurring after sars-cov-2 vaccine. J Blood Med 2021;12:221–4. https://doi.org/10.2147/JBM.S307047.

[30] Burn E, Li X, Delmestri A, Jones N, Duarte-Salles T, Reyes C, Martinez-Hernandez E, Marti E, Verhamme KMC, Rijnbeek PR, Strauss VY, Prieto-Alhambra D. Thrombosis and thrombocytopenia after vaccination against and infection with SARS-CoV-2 in the United Kingdom. Nat Commun 2022;13(1):1–10. https://doi.org/10.1038/s41467-022-34668-w. *2022 13:1*.

[31] Burn E, Roel E, Pistillo A, Fernández-Bertolín S, Aragón M, Raventós B, Reyes C, Verhamme K, Rijnbeek P, Li X, Strauss VY, Prieto-Alhambra D, Duarte-Salles T. Thrombosis and thrombocytopenia after vaccination against and infection with SARS-CoV-2 in Catalonia, Spain. Nat Commun 2022;13(1):1–11. https://doi.org/10.1038/s41467-022-34669-9. *2022 13:1*.

[32] Feese MD, Tamada T, Kato Y, Maeda Y, Hirose M, Matsukura Y, Shigematsu H, Muto T, Matsumoto A, Watarai H, Ogami K, Tahara T, Kato T, Miyazaki H, Kuroki R. Structure of the receptor-binding domain of human thrombopoietin determined by complexation with a neutralizing antibody fragment. Proc Nat Acad Sci USA 2004;101(7):1816–21. https://doi.org/10.1073/pnas.0308530100.

[33] *GitHub - project-gemmi/gemmi: macromolecular crystallography library and utilities*. (n. d.). Retrieved February 16, 2022, from https://github.com/project-gemmi/gemmi.

[34] Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 2019. https://doi.org/10.1093/nar/gkz268.

[35] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Vázquez-Baeza Y. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2. *2020 17:3*.