Structural Bioinformatics

Epitopedia: identifying molecular mimicry of known immune epitopes

Christian A Balbin[1*], Janelle Nunez-Castilla[1] and Jessica Siltberg-Liberles[1,2,*]

[1]Department of Biological Sciences, College of Arts, Science and Education, and [2]Biomolecular Sciences Institute, Florida International University, 11200 SW 8th St, Miami, 33199, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Upon infection, pathogen epitopes stimulate the host's immune system to produce antibodies targeting the pathogen. Molecular mimicry (structural similarity) between an infecting pathogen and host proteins or pathogenic proteins the host has previously encountered can impact the immune response of the host. The ability to identify potential molecular mimicry for a pathogen can illuminate immune effects with importance to pathogen treatment and vaccine design.

Summary: Epitopedia allows for identification of regions with three-dimensional molecular mimicry between a protein in a pathogen with known epitopes in the host.

Results: SARS-CoV-2 Spike returns molecular mimicry with 14 different epitopes including integrin beta-1 from *Homo sapiens*, lethal factor precursor from *Bacillus anthracis*, and pollen allergen Phl p 2 from Timothy grass.

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

Contact: cbalbin@fiu.edu, jliberle@fiu.edu

1 Introduction

Pathogens present antigenic molecules that typically elicit a host immune response. For proteins, an epitope is the portion of the antigen which is bound by an antibody. Occasionally, pathogen epitopes may resemble host epitopes, a phenomenon termed molecular mimicry. In instances of molecular mimicry, infection with a pathogen can trigger the production of antibodies that mistakenly target an epitope in a host protein, resulting in autoimmune disease (Cusick *et al.*, 2012). Alternatively, molecular mimicry between two pathogens can offer protective immunity for both after infection with either one (Agrawal, 2019).

To the best of our knowledge there are currently no computational programs or pipelines readily available for the prediction of molecular mimicry of known epitopes, although programs to map peptides (mimotopes) onto the antigenic protein structure to identify a native epitope exist (Huang *et al.*, 2008; Mayrose *et al.*, 2007; Negi and Braun, 2009; Chen *et al.*, 2012).

We present Epitopedia, a computational pipeline for the prediction of molecular mimicry. Epitopedia identifies sequence and structural similarity between an antigenic protein of interest and any experimentally verified linear epitope found in the Immune Epitope Database (IEDB) (Vita *et al.*, 2019). Given the structural similarity between these epitopes and the pathogenic protein, it follows that binding of the same antibody may be possible.

2 Implementation

2.1 Internal Database Generation

Epitopedia utilizes IEDB and the Protein Data Bank (PDB) (Berman et al., 2000) to generate four internal databases. IEDB-FILT is derived from the IEDB database, which is reduced to only include the necessary data for mimicry discovery including the full-length source sequences. A BLASTP database (EPI-SEQ), including taxonomic origin, is generated from epitope linear peptide sequences (mean length 13 residues) with positive assays in IEDB. A repository of structural representatives (EPI-PDB) for the source sequences is generated from a sensitive (s=7.5) MMseqs2 (Steinegger and Söding, 2017) many-against-many search against the PDB. Alignments with less than 90% identity or 20% query coverage are discarded. Lastly, DSSP (Kabsch and Sander, 1983) is used to compute the accessible surface area (ASA) for every residue in each chain in PDB. The generated databses are then stored as tables in a SQLite3 database.

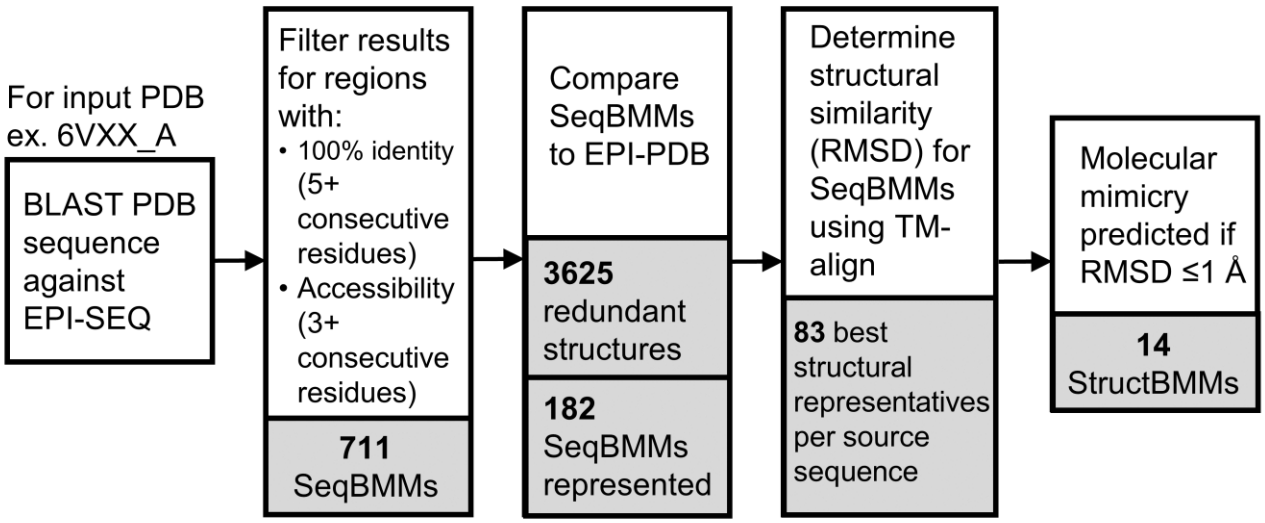

Figure 1. Epitopedia output overview based on input PDB 6VXX, chain A. For example of detailed output see example_output folder on the GitHub repository.

2.2 Searching for Sequence-Based Molecular Mimics

The input for Epitopedia is one or more PDB structures. The protein sequence of the input structure is used to BLAST against EPI-SEQ. The BLASTP parameters evalue and max_target_seq are set to 2,000,000 to avoid discarding hits due to large evalue or reaching the match limit. The hits are filtered to only include hits with regions of 5 consecutive, identical amino acids between the query (input protein) and subject (epitope). If a hit meets this requirement for more than one region, the regions are split into subalignments (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 > 20%. Relative surface area is computed according to equation

$$RSA = ASA/MaxASA$$

with Wilke (Tien *et al.*, 2013) providing the maximum allowed solvent accessibility (MaxASA) per amino acid. Regions meeting these qualifications are considered sequence-based molecular mimics (SeqBMMs); those that do not meet the qualifications are discarded.

2.3 Identifying Structural Molecular Mimics

The structural regions of the input structure corresponding to the SeqBMMs' regions are evaluated to ensure that all residues are solved. To avoid potential problems of missing mimics in an input structure due to e.g., missing electron density, several structures can simultaneously be used as an input, helpful in the case of unsolved regions in some structures and allowing for the representation of a conformational ensemble. SeqBMMs represented in EPI-PDB and the corresponding hit fragment from the input structure are extracted. TM-align (Zhang and Skolnick, 2005) is used to evaluate the structural similarity for each extracted peptide structure pair. The alignment of the identical mimic region for the peptide pair is provided to TM-align which then performs the structural superposition and generates an RMSD value. Pairs with an RMSD ≤ 1Å are considered structural molecular mimics (StructBMMs).

2.4 Handling Redundancy and Visualization

It is common to have several overlapping epitopes where both the StructBMM region and epitope source sequences are identical for multiple SeqBMMs. Internal accession numbers for all epitope source sequences in IEDB-FILT were assigned, such that any two or more identical sequences will have the same internal accession number. This allows 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 and Bootstrap.

3 Results

Executing Epitopedia with SARS-CoV-2 Spike protein (PDB id: 6VXX, chain A (Walls *et al.*, 2020)) as input yields 711 SeqBMMs, where 182 SeqBMMs from 83 source sequences have PDB representation (Figure 1). Based on a cutoff of 1 Å, there are 14 StructBMMs. Of the 14 epitopes with molecular mimicry to Spike, 11 are from human such as integrin beta-1, and one each are from *Mycobacterium tuberculosis*, *Bacillus anthracis*, and Timothy grass (Table S1, Figures S1-S12). However, proteins are dynamic and the input PDB is important for the results. Thus, Epitopedia allows a set of PDB ids representing a conformational ensemble of the same protein or of different proteins to be used as input. If a conformational ensemble of the same protein is used, the pipeline will run for each PDB id, but at the end, all results will be considered in determining the structural molecular mimics based on the RMSD cutoff.

Acknowledgements

We thank the RAPID group at FIU for weekly discussions and Trevor Cickovski for testing the software.

Funding

This work was partially supported by the National Science Foundation under Grant No. 2037374 (JNC and JSL).

Conflict of Interest: none declared.

References

Agrawal,B. (2019) Heterologous Immunity: Role in Natural and Vaccine-Induced Resistance to Infections. *Front. Immunol.*, **0**, 2631.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Chen,W. *et al.* (2012) Pepmapper: A collaborative web tool for mapping epitopes from affinity-selected peptides. *PLoS One*, **7**, 37869.

Cusick,M.F. *et al.* (2012) Molecular mimicry as a mechanism of autoimmune disease. *Clin. Rev. Allergy Immunol.*, **42**, 102–111.

Huang,Y.X. *et al.* (2008) Pep-3D-Search: A method for B-cell epitope prediction based on mimotope analysis. *BMC Bioinformatics*, **9**.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Mayrose,I. *et al.* (2007) Pepitope: Epitope mapping from affinity-selected peptides. *Bioinformatics*, **23**, 3244–3246.

Negi,S.S. and Braun,W. (2009) Automated detection of conformational epitopes using phage display peptide sequences. *Bioinform. Biol. Insights*, **2009**, 71–81.

Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol. 2017 3511*, **35**, 1026–1028.

Tien,M.Z. *et al.* (2013) Maximum allowed solvent accessibilites of residues in proteins. *PLoS One*, **8**.

Vita,R. *et al.* (2019) The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.*, **47**, D339–D343.

Walls,A.C. *et al.* (2020) Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*, **181**, 281-292.e6.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.

Supplementary material

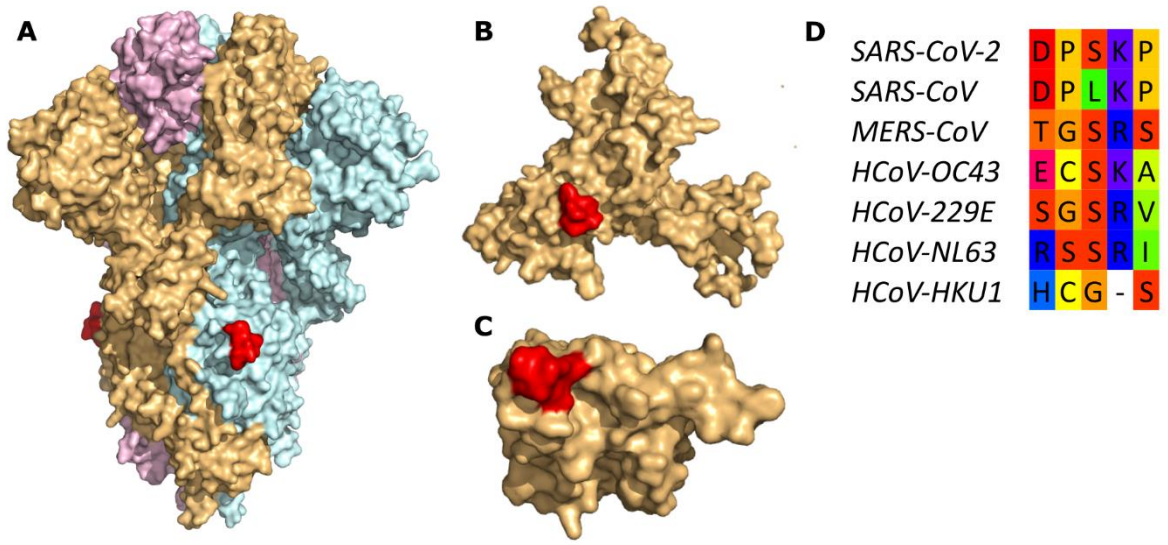

Fig S1. The molecular mimicry motif DPSKP (red) from Spike (A, colored by chain) matches ribosomal protein L3 (B, beige) from *Homo sapiens* with an RMSD of 0.09 Å and alanine and proline-rich secreted protein apa precursor (C, beige) from *Mycobacterium tuberculosis* with an RMSD of 0.22 Å. The motif is not conserved in human betacoronaviruses (D). Protein structures visualized can be found in Table S1. Sequences for human betacoronavirus Spike proteins were aligned using MAFFT. The molecular mimicry motif region was extracted from the alignment according to Table S1. Accessions for the sequences in order of appearance are: YP_009724390, YP_009825051, YP_009047204, YP_009555241, NP_073551, YP_003767, YP_173238.

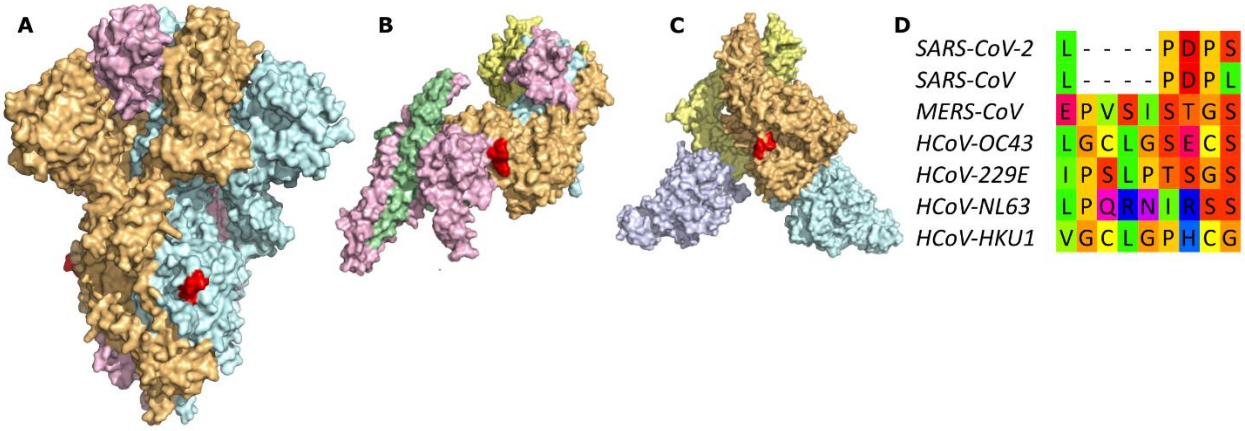

Fig S2. The molecular mimicry motif LPDPS (red) from Spike (A, colored by chain) matches BRCA1-A complex subunit BRE (B, colored by chain) from *Homo sapiens* with an RMSD of 0.18 Å and semaphoring-7A (C, colored by chain) from *Homo sapiens* with an RMSD of 0.66 Å. The motif is not conserved in human betacoronaviruses (D). For details, see legend of Fig S1.

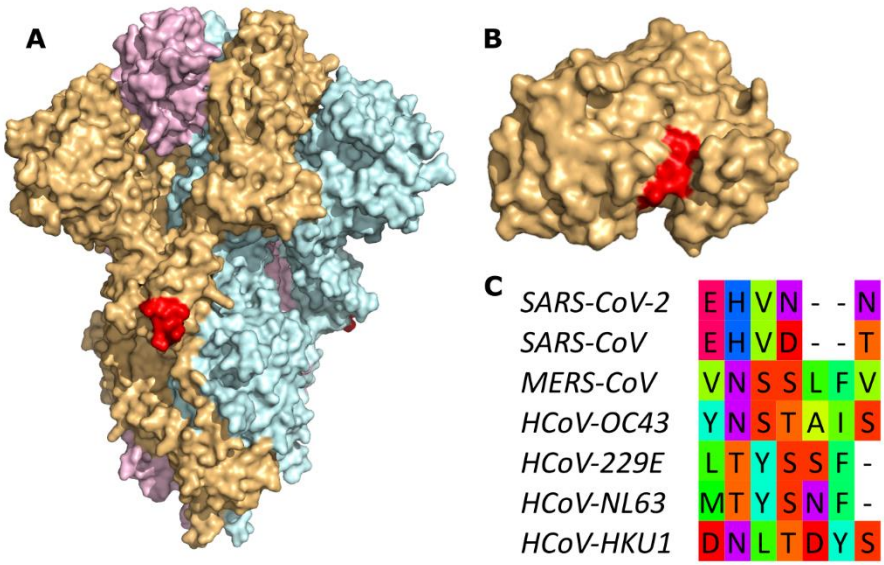

Fig S3. The molecular mimicry motif EHVNN (red) from Spike (A, colored by chain) matches casein kinase 2 alpha isoform (B, beige) from *Homo sapiens* with an RMSD of 0.30 Å. The motif is not conserved in human betacoronaviruses (C). For details, see legend of Fig S1.

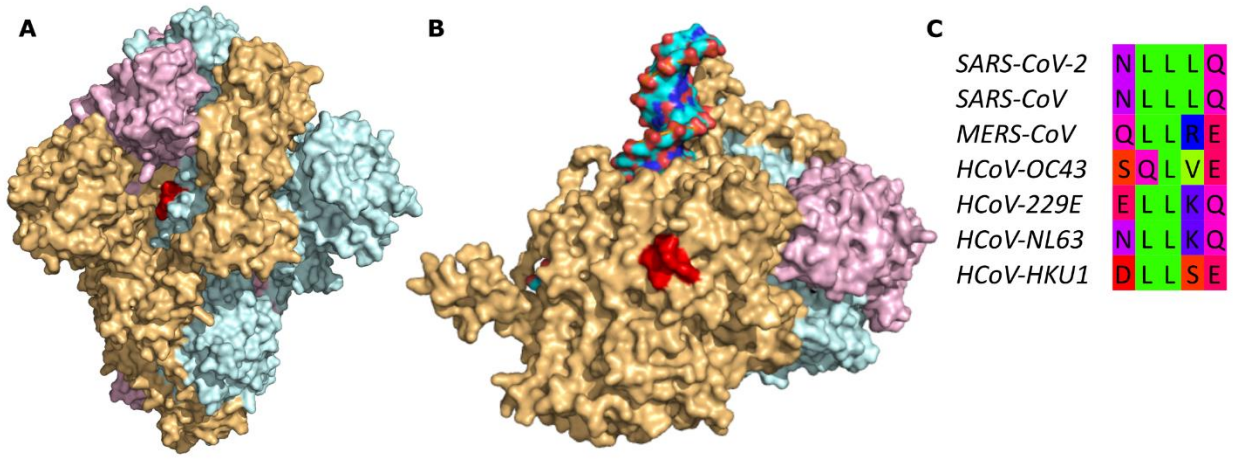

Fig S4. The molecular mimicry motif NLLLQ (red) from Spike (A, colored by chain) matches DNA polymerase subunit gamma-1 (B, colored by chain, with DNA colored by element) from *Homo sapiens* with an RMSD of 0.42 Å. The motif is semi-conserved in human betacoronaviruses (C). For details, see legend of Fig S1.

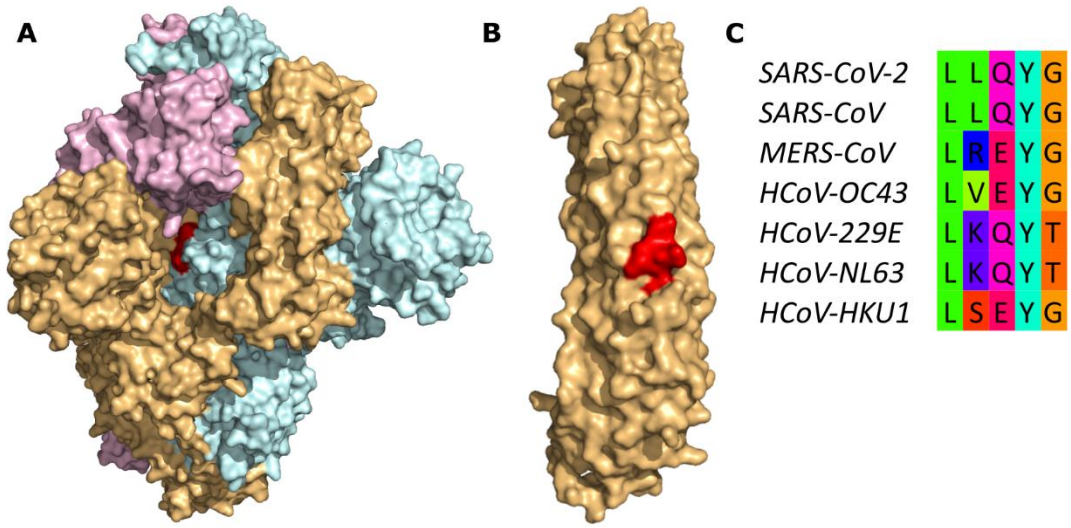

Fig S5. The molecular mimicry motif LLQYG (red) from Spike (A, colored by chain) matches ankyrin-1 (B, beige) from *Homo sapiens* with an RMSD of 0.49 Å. The motif is semi-conserved in human betacoronaviruses (C). For details, see legend of Fig S1.

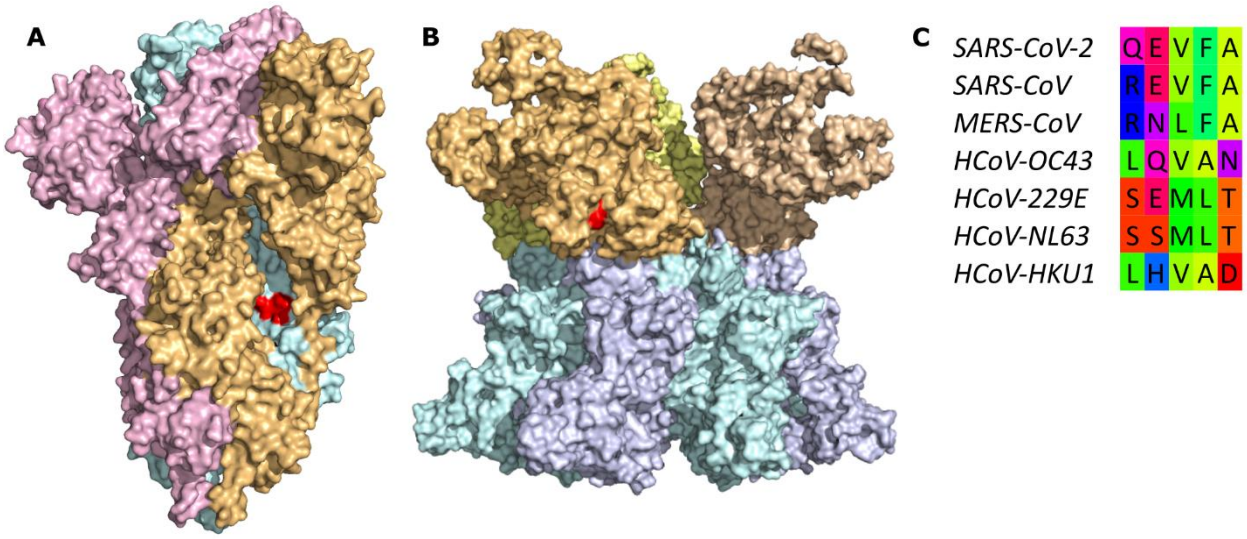

Fig S6. The molecular mimicry motif QEVFA (red) from Spike (A, colored by chain) matches lethal factor precursor (B, colored by chain) from *Bacillus anthracis* with an RMSD of 0.59 Å. The motif is not conserved in human betacoronaviruses (C). For details, see legend of Fig S1.

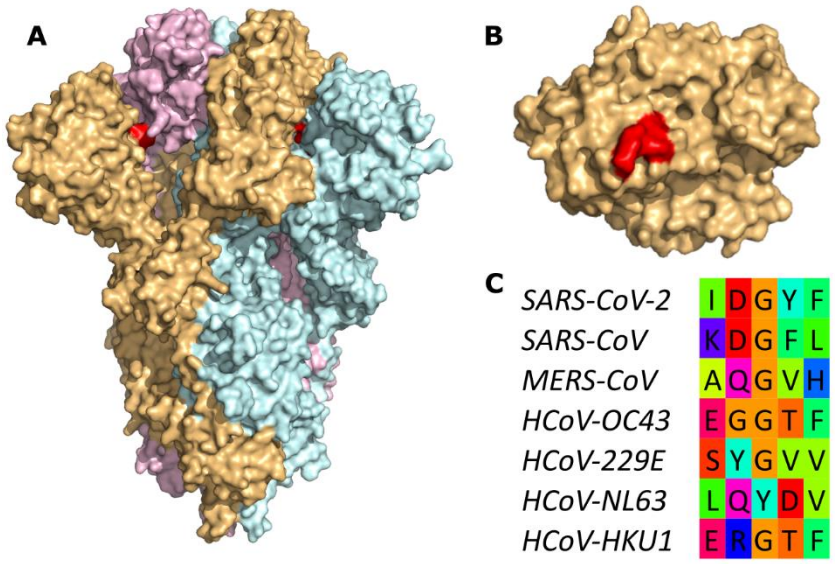

Fig S7. The molecular mimicry motif IDGYF (red) from Spike (A, colored by chain) matches lanosterol 14-alpha demethylase (B, beige) from *Homo sapiens* with an RMSD of 0.64 Å. The motif has low conservation in human betacoronaviruses (C). For details, see legend of Fig S1.

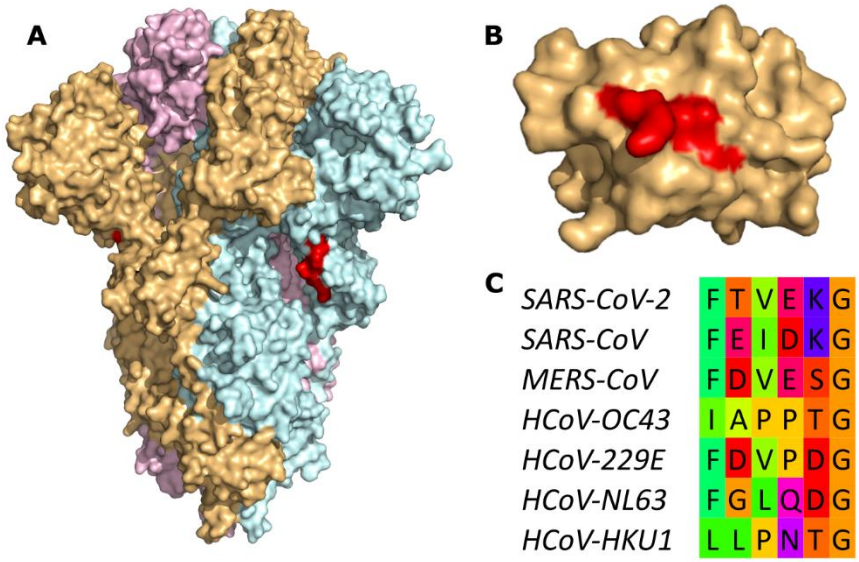

Fig S8. The molecular mimicry motif FTVEKG (red) from Spike (A, colored by chain) matches pollen allergen Phl p 2 (B, beige) from *Phleum pratense* with an RMSD of 0.67 Å. The motif is semi-conserved in human betacoronaviruses (C). For details, see legend of Fig S1.

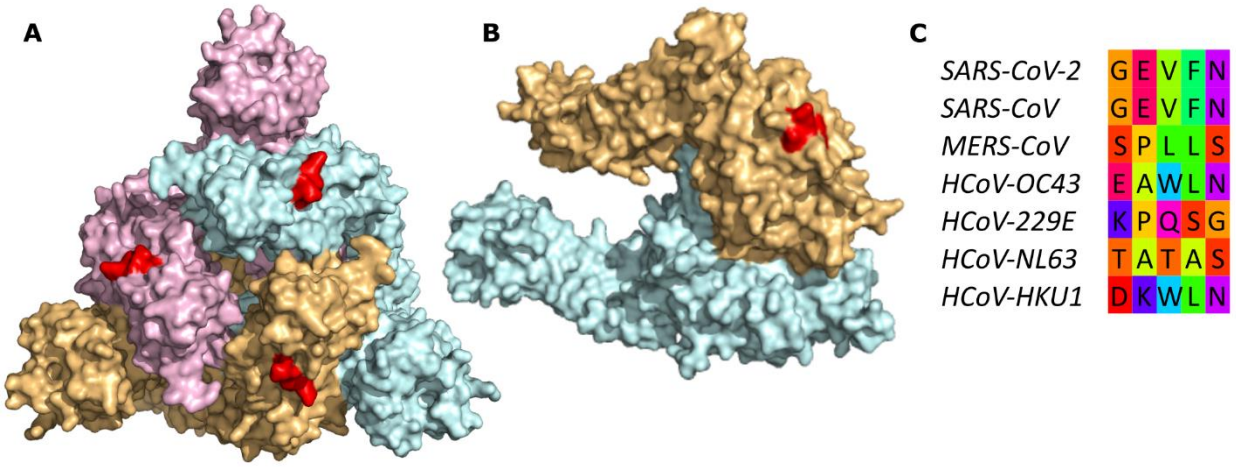

Fig S9. The molecular mimicry motif GEVFN (red) from Spike (A, colored by chain) matches integrin beta-1 (B, colored by chain) from *Homo sapiens* with an RMSD of 0.67 Å. The motif is not conserved in human betacoronaviruses (C). For details, see legend of Fig S1.

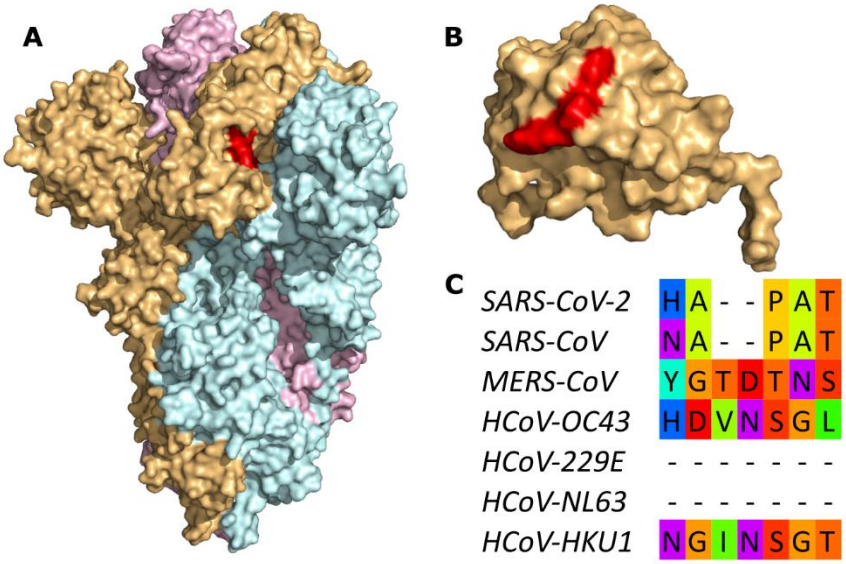

Fig S10. The molecular mimicry motif HAPAT (red) from Spike (A, colored by chain) matches activator of 90 kDa heat shock protein ATPase homolog 1 (B, beige) from *Homo sapiens* with an RMSD of 0.76 Å. The motif is not conserved in human betacoronaviruses (C). For details, see legend of Fig S1.

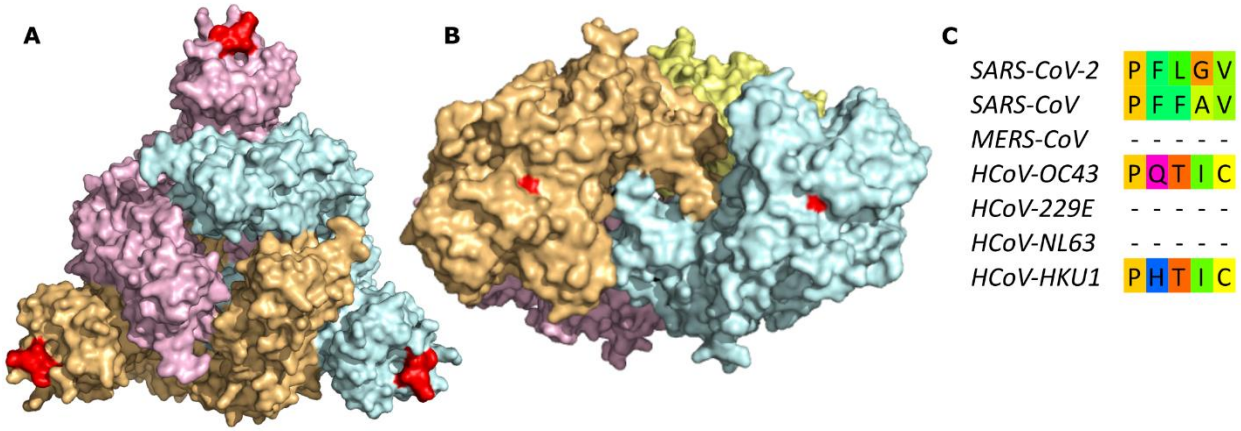

Fig S11. The molecular mimicry motif PFLGV (red) from Spike (A, colored by chain) CTP synthase 1 (B, colored by chain) from *Homo sapiens* with an RMSD of 0.77 Å. The motif is not conserved in human betacoronaviruses (C). For details, see legend of Fig S1.

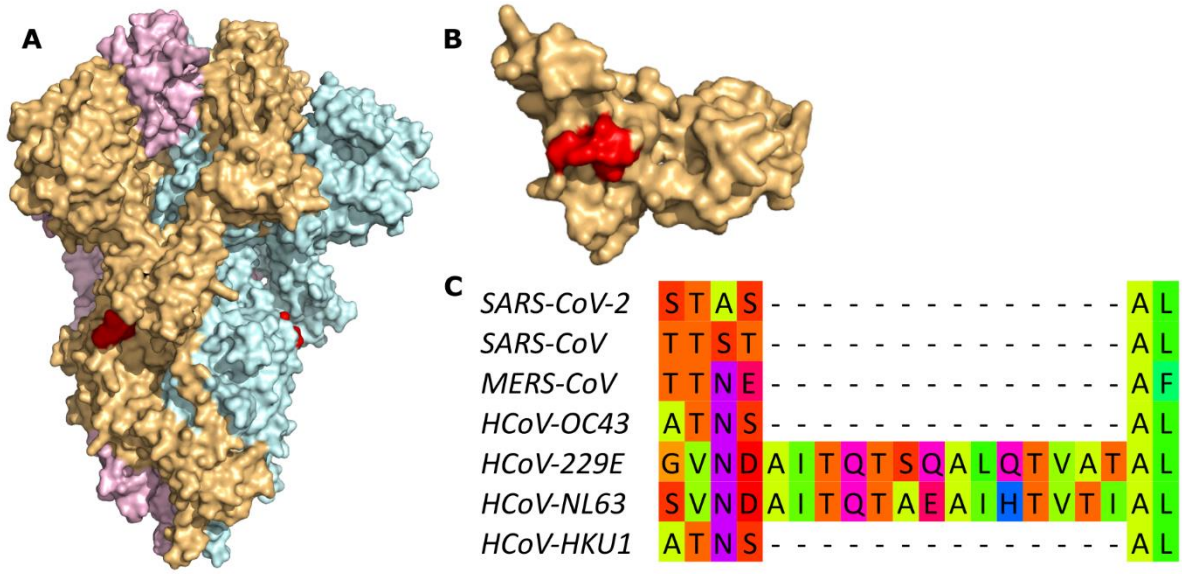

Fig S12. The molecular mimicry motif STASAL (red) from Spike (A, colored by chain) 40S ribosomal protein S13 (B, beige) from *Homo sapiens* with an RMSD of 0.84 Å. The motif is not conserved in human betacoronaviruses (C). For details, see legend of Fig S1.