Supplementary Material

PESI: Paratope-Epitope Set Interaction for SARS-CoV-2 Neutralization Prediction

1. DATASET CURATION

Following the same numbering and delineated epitope sites as in [1], we mapped the epitope residues onto the full 3D structure of the SARS-CoV2 S protein (PDB code 7VXF) [2]. We also align the epitope sites against the reference strain sequence hCoV - 19/Wuhan/W IV 04/2019 (WI04) provided by the GISAID database [3] with accession number EPI_ISL_402124. (WI04). Table S1 shows the epitope sites' residue positions and their corresponding sequence in WI04.

We then downloaded the crystal structures of each antibody-antigen complex from PDB in cif format. We also isolated columns which are relevant for this work, namely antibody name and neutralization class against SARS-CoV-2. The antibody-antigen complex (.cif) files were then visualized with PyMOL [4]. While most contained an antibody in complex with the SARS-CoV2 S protein, some contained an antibody in complex with other viruses. As the co-crystal structures were often in trimeric complex form, a single copy of the antibody-antigen complex in correct orientation was isolated by manual inspection and selection of chains. The paratope region of the antibody was isolated by filtering amino acid residues within $5\mathring{A}$ of the antigen S protein. The SARS-CoV2 reference S protein (PDB 7VXF) was then loaded as an additional object in the same PyMOL instance and aligned with the respective file's S protein. The epitope region on the reference S protein was identified by filtering amino acid residues within $5\mathring{A}$ of the paratope. Thereafter, the epitope and paratope sequences were extracted for each antibody-antigen complex and added to the database. The epitope sequences were cross-matched with the epitope candidates from Table S1 [1] to ensure correct placement.

We directly concatenate the CDR fragments for the paratope residues and the nine epitope sites for the epitope residues In total, 310 paratope-epitope pairs are obtained, among which 228 sequence pairs are neutralizing (labeled as 1) while the remaining 82 samples are non-neutralizing (labeled as 0).

Our approach of delineating epitope-paratope pairs is a merger of linear and conformational epitopes and can be considered as interpolated epitope and paratope pairs, respectively. An advantage of our modeling approach is that it allows the detection of neutralization from epitope or paratope sequences alone. Thereby it could be suitable for practical applications in rapid, high-throughput in-silico identification from a large list of candidate epitopes or paratopes, such as based on motif-scans of antibody repertoires in the initial stages. It can be complementary with later pipelines of extensive modeling to assess antibody binding affinity such as in those required by scaffolding onto framework regions, structure re-construction etc.

REFERENCES

1. M. Sikora, S. von Bülow, F. E. Blanc, M. Gecht, R. Covino, and G. Hummer, "Computational epitope map of sars-cov-2 spike protein," PLoS computational biology **17**, e1008790 (2021).
2. Y. Wang, C. Xu, Y. Wang, Q. Hong, C. Zhang, Z. Li, S. Xu, Q. Zuo, C. Liu, Z. Huang *et al.*, "Conformational dynamics of the beta and kappa sars-cov-2 spike proteins and their complexes with ace2 receptor revealed by cryo-em," Nat. communications **12**, 7345 (2021).
3. O. Re3data, "Global Initiative on Sharing Avian Influenza Data," (2012). Freunde von GISAID e.V. Germany. Registry of Research Data Depositories (re3data.org). Global Initiative on Sharing Avian Influenza Data, 2012. https://gisaid.org/. Accessed Feb 2022.
4. L. Schrödinger, "The pymol molecular graphics system," .

Table S1. Epitope Candidates on SARS-CoV2 S Protein. Numbering corresponds to [1], using reference structure with PDB ID 7VXF. Sequence corresponds to reference strain WI04.

Epitope	Residues	Sequence
E1	15-28	CVNLTTRTQLPPAY
	63-79	TWFHAIHVSGTNGTKRF
	247-260	SYLTPGDSSSGWTA
E2	97	K
	178-189	DLEGKQGNFKNL
	207-219	HTPINLVRDLPQG
E3	137-164	NDPFLGVYYHKNNKSWMESEFRVYSSAN
E4	332-346	ITNLCPFGEVFNATR
E5	403-406	RGDE
	438	S
	440-451	NLDSKVGGNYNY
	495-506	YGFQPTNGVGYQ
E6	452-476	LYRLFRKSNLKPFERDISTEIYQAG
	479-482	PCNG
	484-494	EGFNCYFPLQS
E7	527-537	PKKSTNLVKNK
E8	603-605	NTS
	633-642	WRVYSTGSNV
	656-661	VNNSYE
	674-693	
E9	808-814	DPSKPSK