

Supplementary Materials for

Identification and validation of 174 COVID-19 vaccine candidate epitopes reveals low performance of common epitope prediction tools

Marek Prachar^{1,2,3}, Sune Justesen³, Daniel Bisgaard Steen-Jensen³, Stephan Thorgrimsen³, Erik Jurgons⁴, Ole Winther^{1,2,5}, and Frederik Otzen Bagger^{1,6,7,*}

1. Center for Genomic Medicine, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark
2. Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark
3. Immunitrack ApS, Copenhagen, Denmark
4. INTAVIS Peptide Services GmbH & Co.KG, Waldhäuser Str. 64, 72076 Tübingen, Germany
5. Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
6. UKBB Universitäts-Kinderspital, Department of Biomedicine, Basel, 4031 Basel, Switzerland
7. Swiss Institute of Bioinformatics, Basel, 4053 Basel, Switzerland

* Correspondence to Frederik Otzen Bagger, frederik.otzen.bagger@regionh.dk

This PDF file includes:

Supplementary methods

Figure S1

Table S1

Table S2

Table S3

Captions for Data S1-S2

Supplementary methods

Analysis of predictions

To produce the ROC graphs and to calculate their AUC we used RStudio Version 1.2.5033, R version 3.6.2, pROC 1.16.1 package. The same software was used to calculate and plot Spearman correlations, using the ggplot2 3.2.1 package.

PrdX

The software was written in Python Version 3.7.3, NN architecture was built using PyTorch Version 1.3.0. For training we used a learning rate of 0.003 and used the Adam optimizer.

ROC metrics

$$Sn = \frac{TP}{TP + FN} = 1 - \frac{FN}{P} \quad 0 \leq Sn \leq 1$$

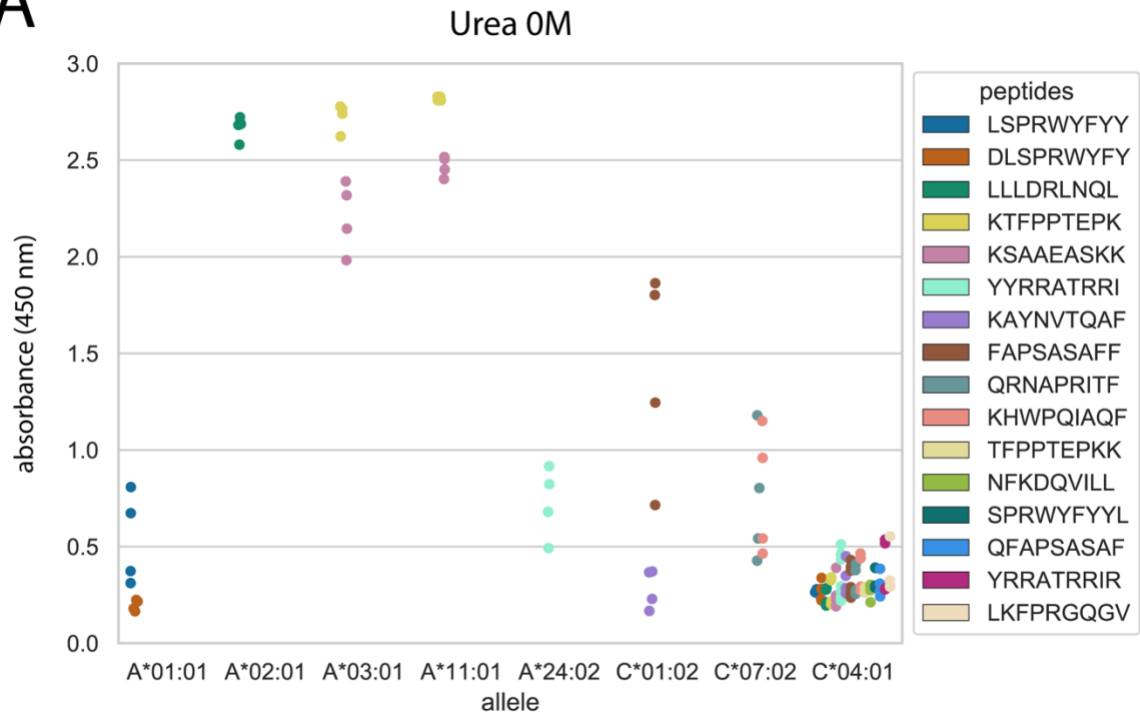
$$Sp = \frac{TN}{TN + FP} = 1 - \frac{FP}{N} \quad 0 \leq Sp \leq 1$$

$$FPR = \frac{FP}{FP + TN} = 1 - Sp \quad 0 \leq FPR \leq 1$$

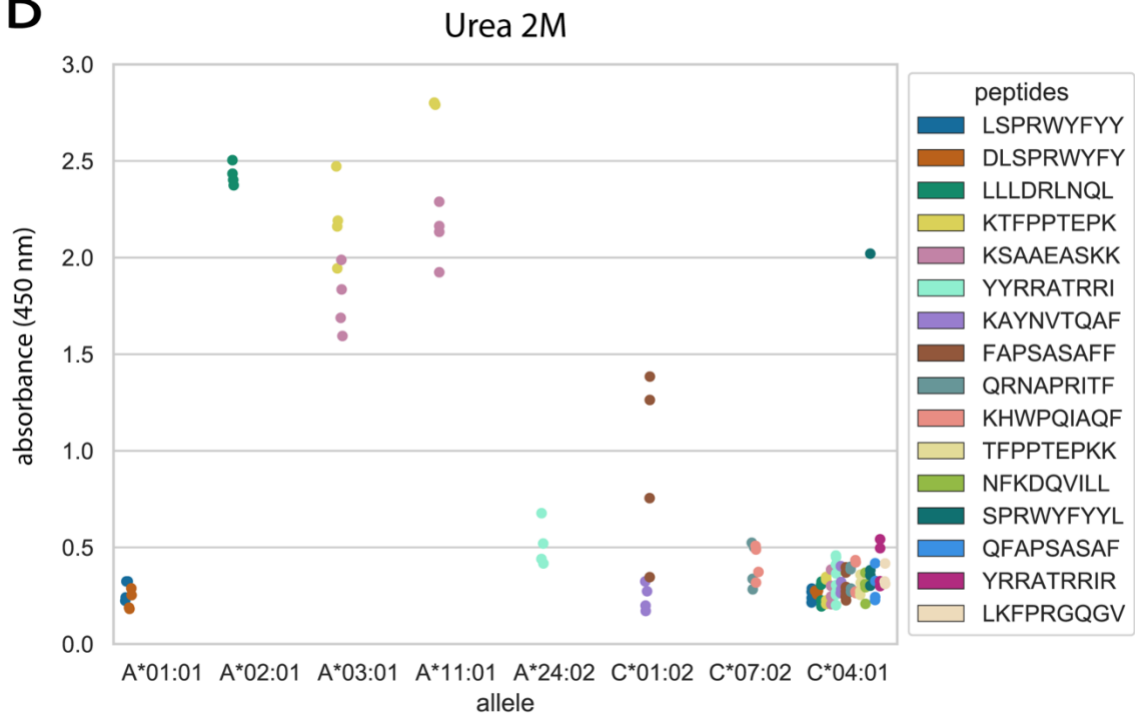
$$Pr = \frac{TP}{TP + FP} \quad 0 \leq Pr \leq 1$$

where Sn, Sp, FNR and Pr stands for sensitivity (recall), specificity, false negative rate and precision, respectively. P, FN, N and FP represent the numbers of positives, false negatives, negatives and false positives, respectively. Rates are converted into percentages in the plots in the manuscript.

A



B



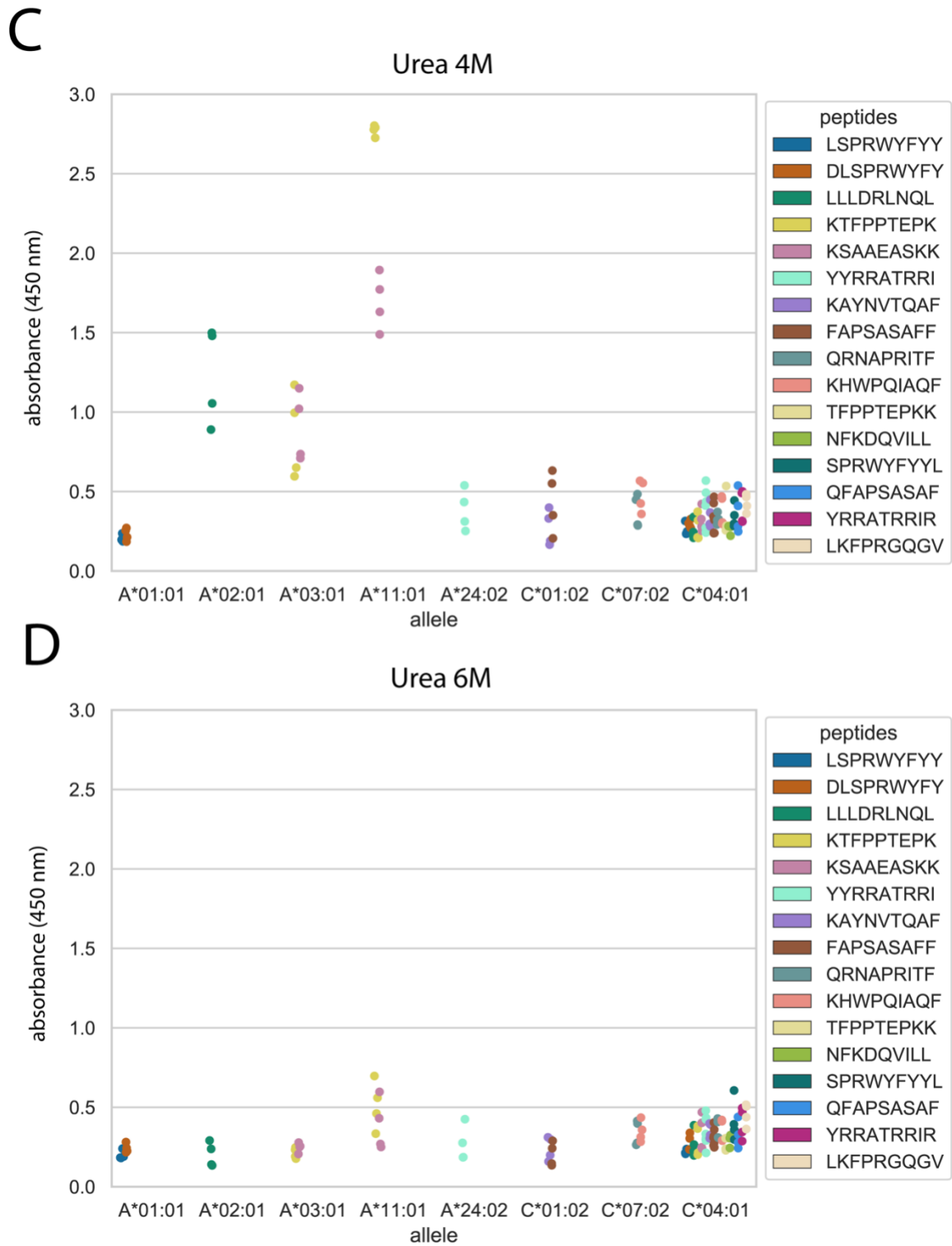


Figure S1. NeoScreen stability assay replicate analysis (n=4), each dot represents a measurement, colored by peptide. Some peptides were measured for multiple alleles. A), B), C), D) correspond to different urea molarity - 0, 2, 4, 6, respectively. Most measurements fall close together, with an average standard deviation of 0.10 and with an average mean of 0.56 between replicates. Majority of the experiments were performed on allele HLA-C*04:01.

Table S1. The prevalence across four different ethnic groups of the selected HLA class I alleles.¹

<i>Allele</i>	<i>Caucasian (%)</i>	<i>Black (%)</i>	<i>Hispanic (%)</i>	<i>Asian or Pacific Islander (%)</i>
<i>A*01:01</i>	27	10	11	5
<i>A*02:01</i>	45	22	37	18
<i>A*03:01</i>	23	18	14	3
<i>A*11:01</i>	15	2	11	38
<i>A*24:02</i>	12	2	24	33
<i>B*40:01</i>	13	1	2	15
<i>C*01:02</i>	5	1	9	27
<i>C*04:01</i>	21	29	25	14
<i>C*07:01</i>	27	25	20	1
<i>C*07:02</i>	21	18	24	33

Table S2. Benchmark comparison, AUC and SCC percentile ranks were calculated per benchmark performance of each included tool. Then the average of AUC and SCC was calculated.²

<i>Tool \ Benchmark</i>	<i>IEDB (%)²</i>	<i>Current study (%)</i>
MHCflurry 1.2.0/1.3.0	80	63
NetMHCpan 4.0	70	67
NetMHC 4.0	70	72
NetMHCcons 1.1	60	67
SMMPMBEC 1.0	47	46
IEDB-AR Consensus	39	77
PickPocket 1.1	38	34
SMM 1.0	37	52

Table S3. List of reference peptides used for their corresponding allele.

HLA Allele	Sequence	Reference
<i>A*01:01</i>	VTEHDTLLY	http://www.iedb.org/epitope/71290
<i>A*02:01</i>	VLDFAPPGA	Wilms tumor antigen 1
<i>A*03:01</i>	AVAHKVHLMYK	https://www.iedb.org/epitope/419554
<i>A*11:01</i>	AVFDRKSDAK	https://www.iedb.org/epitope/5316
<i>A*24:02</i>	AYAQKIFKIL	https://www.iedb.org/epitope/5731
<i>B*40:01</i>	REDQWCGSL	https://www.iedb.org/epitope/53476
<i>C*01:02</i>	QYDPVAALF	http://www.iedb.org/epitope/52886
<i>C*04:01</i>	QYDPVAALF	http://www.iedb.org/epitope/52886
<i>C*07:01</i>	YLHARLREL	Identified by previous screening
<i>C*07:02</i>	NYFNRMFHF	Identified by previous screening
<i>DRB1*04:01</i>	AKFVAAWTLKAAA	https://www.iedb.org/epitope/2192

Data S1. COVID-19 INTAVIS-Immunitrack dataset

<https://www.immunitrack.com/wp/wp-content/uploads/Covid19-Intavis-Immunitrack-datasetV2.xlsx>

Excel file containing results of the NeoScreen assay.

Data S2. Full list of predicted and measured peptides used in the benchmark

<https://doi.org/10.5281/zenodo.4277689>

Public datasets of all peptides that were predicted by the NetMHC suite tools and subsequently measured. List of peptide sequences for each respective allele. All predictions for each of the benchmarked tools (not all tested tools use the same measure). List of binders (epitopes) with IEDB links and T cell studies overlap.

References

1. Gonzalez-Galarza, F. F. *et al.* Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.* (2019) <https://doi.org/10.1093/nar/gkz1029>.
2. Trolle, T. *et al.* Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* **31**, 2174–2181 (2015).