PREDICTING THE BINDING OF SARS-COV-2 PEPTIDES TO THE MAJOR HISTOCOMPATIBILITY COMPLEX WITH RECURRENT NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting the binding of viral peptides to the major histocompatibility complex with machine learning can potentially extend the computational immunology toolkit for vaccine development, and serve as a key component in the fight against a pandemic. In this work, we adapt and extend *USMPep*, a recently proposed, conceptually simple prediction algorithm based on recurrent neural networks. Most notably, we combine regressors (binding affinity data) and classifiers (mass spectrometry data) from qualitatively different data sources to obtain a more comprehensive prediction tool. We evaluate the performance on a recently released SARS-CoV-2 dataset with binding stability measurements. *USMPep* not only sets new benchmarks on selected single alleles, but consistently turns out to be among the best-performing methods or, for some metrics, to be even the overall best-performing method for this task.

1 INTRODUCTION

Predicting the binding between viral peptides and human proteins from the adaptive immune system using machine learning may serve as a valuable tool to increase the speed of vaccine development in the ongoing SARS-CoV-2-pandemic as well as in future health crises. Accelerated vaccine development supported by computational biology tools may become especially relevant against the background of an evolutionary arms race between viral escape variants and vaccine adaptation until herd immunity can finally be reached.

Major histocompatibility complex (MHC) molecules encoded by the human leukocyte antigen (HLA) gene complex, play a crucial role in the adaptive immune system (Klein & Sato, 2000; Wieczorek et al., 2017). They induce an immune response by presenting antigen fragments on the cell-surface to immune effector cells (Wieczorek et al., 2017; Vyas et al., 2008) , and therefore take part in gaining acquired immunity through vaccination. E.g., novel RNA-based vaccines against SARS-CoV-2 enter human cells and elicit the expression of viral spike proteins. They are broken down by the proteasome into antigen peptides which bind to MHC proteins with varying binding affinity. Bound antigen peptides (protein-derived epitopes) are presented by MHC on the cell surface and tie to T-cells that trigger an immune response leading to acquired immunity (Sahin et al., 2014).

MHC is highly polymorphic such that humans express individual combinations of MHC alleles that bind differently tight to a given peptide, which can affect the potency of an evoked immune response (Winchester, 2008). Moreover, there are different MHC classes. MHC class I molecules are found on almost every nucleated body cell and on platelets at varying densities. They continuously present fragments of proteins produced in the cell – self or non-self antigens (e.g., viruses) – to CD8 T cells (Groothuis et al., 2005; Shastri et al., 2005). MHC class II occurs mainly in professional antigen presenting cells of the immune system (e.g., B-lymphocytes) where they present fragments of extracellular ingested pathogens to CD4 T cells (Vyas et al., 2008).

At present, several (e.g., mRNA-based) COVID-19-vaccines make use of the amino acid sequence of the SARS-CoV-2 spike protein, which constitutes about 1/8 of the viral proteome (Prachar et al., 2020). Viral escape variants of the spike protein that would degrade into peptides that bind less tight to MHC can be expected to become more prevalent due to evolutionary pressure as a result of widespread vaccine campaigns (Weisblum et al., 2020). In this case, it might be necessary to lever-

age selected parts of the remaining 7/8 of the viral proteome for novel vaccine candidates (Prachar et al., 2020; Grifoni et al., 2020). Identifying and increasing the number of immunodominant B- and T-cell epitopes (while excluding those that may even cause adverse effects) is a potential strategy in vaccination design to generate protective immunogenicity (Dong et al., 2020). Multi-epitope vaccines against SARS-CoV-2 might be able to achieve a more precise immune response and to limit the risk of allergic reactions (see Kar et al., 2020). While full experimental characterization of all potential peptides of several virus variants is slow or might not be feasible at all, prioritization by MHC-peptide binding stability prediction may substantially accelerate the development of a more effective vaccine (Prachar et al., 2020; Grifoni et al., 2020). This approach may also enable the creation of epitope vaccines targeted against several virus strains at the same time.

A wide range of binding affinity prediction methods based on machine learning has been developed with potential application to vaccine development as well as to personalized cancer immunotherapy. These methods are summarized in a recent comparative review (Zhao & Sher, 2018); see also Prachar et al. (2020) for a comparison with particular focus on SARS-CoV-2. At this point, it is worth stressing that many of the established methods rely on complicated training procedures with intricate model selection procedures and/or rely on heuristics to identify, e.g., binding regions.

In this work, we evaluate the performance of a novel algorithm for peptide-MHC binding affinity/stability prediction on a recently released dataset with binding stability measurements between SARS-CoV-2 peptides and ten alleles of MHC class I and one allele of MHC class II (Prachar et al., 2020). The algorithm is based on recurrent neural networks and was recently proposed as *USMPep* (Vielhaben et al., 2020). The publication of the dataset by Prachar et al. (2020) also contains a benchmark comparison of about twenty state-of-the-art prediction algorithms (published before 2 March 2020) on these new binding stability measurements. With this contribution, we provide an update for this benchmark by adding the results of an extended version of *USMPep* (which was published on 2 July 2020, i.e. after the 'reporting date' of Prachar et al. (2020)).

2 Materials and Methods

Datasets & Targets Objective of our work is to predict the binding stability between SARS-CoV-2 peptides and MHC based on the amino acid sequences of the peptides. For this purpose, we train and finally evaluate recurrent neural networks on three different types of lab measurements, involving a peptide of known amino acid sequence and a given MHC allele. Therefore, we distinguish three qualitatively different kinds of targets: During training, we encounter binding affinity (BA) for peptides and mass-spectrometry-eluted (MS) ligands. Whereas the former represents a continuous target (leaving aside qualitative binding affinity labels as provided by O'Donnell et al. (2018)), the latter only yield positive (i.e. binding) samples, which are typically combined with artificial negative samples in order to be able to train a classifier on this data. Finally, during test time, we aim to predict binding stability (BS), which is also a continuous target, but quantifies the stability of the binding and is hence a more specific measure than binding affinity (Harndahl et al., 2012; Jørgensen et al., 2014). Due to a lack of appropriate training data, we use BA as a proxy target for BS.

We use a MHC class I BA dataset provided by O'Donnell et al. (2018) and a MS dataset compiled by Jurtz et al. (2017) for model training. For MHC class II alleles, we train our models on a dataset from Jensen et al. (2018), which includes BA and MS data. All datasets are based on data retrieved from the Immune Epitope Database (Vita et al., 2018). MS datasets additionally include artificial decoys. We evaluate our tools on the aforementioned BS dataset provided by Prachar et al. (2020), where the stability measurements are normalized to an allele-specific reference peptide.

Evaluation metrics We consider the most predominantly used metrics in the field (Zhao & Sher, 2018; Prachar et al., 2020), namely Spearman's ρ and the area under the receiver operating curve (AUCROC) upon framing the task as a classification task using a threshold value of 60% stability. In order to compare the overall performance, we follow Vielhaben et al. (2020) and consider summary metrics, in this case the median due to the small number of alleles under consideration, across alleles.

Model We build our approach on *USMPep*, a recently proposed, conceptually simple yet very powerful method (Vielhaben et al., 2020), which is based on a recurrent neural network, in this case with a single-layer long short-term memory (LSTM) architecture. We focus on single-allele

models, and precondition the model weights based on an (up to the classification head) identical architecture pretrained with an autoregressive language model objective (Vielhaben et al., 2020), which generally lead to slight but consistent improvements compared to training from scratch. We consider ensembles of ten individual models for improved stability.

The quantitative and qualitative subsets of the available data let us consider two training objectives: On the log-transformed BA data, we train a regression model (*USMPep_BA*), as in Vielhaben et al. (2020) using a modified mean squared error loss function that allows to include also qualitative BA measurements (O'Donnell et al., 2018). To leverage the additional, complementary data available through qualitative MS measurements, we train separate classification models using the epitopes identified via MS as well as the artificial negative samples provided in the original MS data using a crossentropy loss. Finally, we consider combined BA+MS ensembles (*USMPep_BAMS*) by averaging log-transformed BA and MS predictions, for the first time in the MHC binding prediction literature, to the best of our knowledge. The source code for training our models will be made available.

3 RESULTS

Figure 1 and Table 1 compare the performance of *USMPep* to other state-of-the-art methods. We show the performance on single alleles based on AUCROC in Figure 1. Both *USMPep*-variants improve the current state-of-the-art for allele A*01:01, the allele with the overall best performance. For B*40:01, *USMPep_BA* raises the current state-of-the-art to a new level. *USMPep_BA* is the only tool in the benchmark that is trained on BA data and achieves the highest AUCROC on more than one allele. While *USMPep* is one of the few tools that provide predictions for the only MHC class II allele in the test set (DRB1*04:01), its performance on this allele is weaker in comparison to the few other available tools.

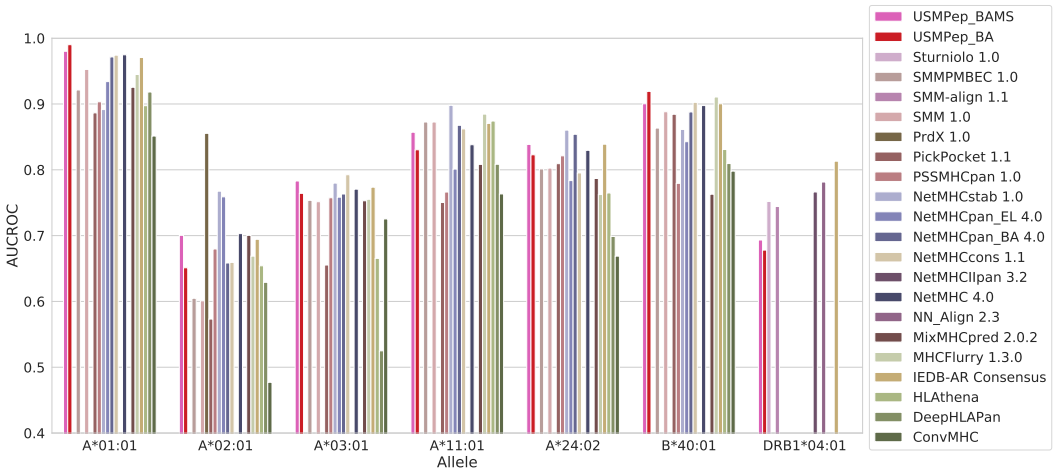

Figure 1: Predictive performance on single alleles: Performance of *USMPep* in predicting the binding stability of SARS-CoV-2 peptides and MHC alleles in comparison with the state-of-the-art tools reported in Table 2 of Prachar et al. (2020). The prediction problem was framed as classification task, and the predictive performance was measured using AUCROC as metric. Allele DRB1*04:01 belongs to MHC class II, all other alleles to class I.

Turning to the overall predictive performance in terms of median AUCROC and Spearman's ρ as shown in Table 1, *USMPep_BAMS* turns out to be the overall best-performing method among all tools in terms of Spearman's ρ and show the fourth best performance in terms of AUCROC. In particular, the ensembling of five regressors trained on BA data and five classifiers trained on MS data considerably improves the overall performance compared to an ensemble of ten regressors on BA-data alone (*USMPep_BA*). It is noteworthy that the *USMPep_BAMS*-ensemble profits from the diversity of the different predictors, e.g. a regressor trained on both BA and MS data along the lines of (O'Donnell et al., 2018) in an ensemble with a classifier trained on MS data yields a weaker

Table 1: Overall predictive performance: The performance of *USMPep* was assessed with the median Spearman's ρ between predicted binding probability and actual BS across alleles. Besides, the median AUCROC across alleles was evaluated. The results of the state-of-the-art-methods were extracted from Figure 2 of Prachar et al. (2020). Because numerous tools do not provide predictions for alleles C*01:02, C*07:01 and DRB1*04:01, these were excluded for the median of Spearman's ρ and AUCROC. AUCROC was only evaluated on alleles with more than ten stable binders, which further excludes two remaining HLA-C alleles. The five highest scores are marked in bold for both metrics and are underlined for the best-performing methods.

Model	Spearman's ρ	AUCROC
USMPep_BAMS	**0.56085**	**0.84785**
NetMHCstab 1.0	**0.51745**	**0.86080**
NetMHCpan_BA 4.0	**0.51610**	**0.86080**
IEDB-AR Consensus	**0.51440**	**0.85470**
USMPep_BA	**0.50280**	0.82660
NetMHC 4.0	0.49545	0.83385
NetMHCpan_EL 4.0	0.49395	0.79235
NetMHCcons 1.1	0.49285	0.82865
MixMHCpred 2.0.2	0.48000	0.77485
SMMPMBEC 1.0	0.46845	0.83235
SMM 1.0	0.46540	**0.83760**
PSSMHCpan 1.0	0.45220	0.77280
MHCFlurry 1.3.0	0.44265	0.82350
PickPocket 1.1	0.41285	0.77980
ConvMHC	0.38750	0.74430
HLAthena	0.38160	0.79780
DeepHLAPan	0.31720	0.75355

performance (Spearman's ρ 0.5260 and AUCROC 0.8097). In summary, these results establish both the original *USMPep_BA* but in particular the newly proposed *USMPep_BAMS* as strong predictors for MHC binding stability compared to other state-of-the-art tools.

4 Summary and Discussion

We evaluate a novel MHC binding prediction tool on recently published BS measurements involving SARS-CoV-2 peptides. The *USMPep* algorithm is characterized by a conceptually simple architecture and training procedure, can process peptides of arbitrary length and does not rely on further heuristics. In order to exploit more training data, we adapt and extend the algorithm to consider not only quantitative BA, but also qualitative MS measurements. We find a very high overall performance of *USMPep* in comparison to other state-of the-art methods, and *USMPep* even outperforms all existing methods on selected single alleles. The method can potentially extend the computational immunology toolkit, and help to accelerate vaccine development, and to prevent future epidemics.

Several limits of the work should be considered. Training a model (on BA and MS measurements as proxy) in order to predict the binding of a given peptide to a certain MHC allele can only serve as first step. It neither necessarily implies BS (as pointed out by Prachar et al., 2020) nor immunogenicity, nor efficacy, nor safety of a potential (e.g., RNA-based) epitope vaccine derived from the amino acid sequence of the peptide. Moreover, additional BS measurements covering a wider range of MHC alleles appear necessary to realise the full potential of this and other prediction tools; in particular in order to warrant that the global population can profit to the same degree in a fair manner, since MHC allele expression may vary with sex and ethnicity (Schneider-Hohendorf et al., 2018; Quiñones-Parra et al., 2014).

REFERENCES

Yetian Dong, Tong Dai, Yujun Wei, Long Zhang, Min Zheng, and Fangfang Zhou. A systematic review of SARS-CoV-2 vaccine candidates. *Signal transduction and targeted therapy*, 5(1):1–14, 2020. URL https://doi.org/10.1038/s41392-020-00352-y.

Alba Grifoni, John Sidney, Yun Zhang, Richard H Scheuermann, Bjoern Peters, and Alessandro Sette. A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell host & microbe*, 27(4):671–680, 2020. URL https://doi.org/10.1016/j.chom.2020.03.002.

Tom A. M. Groothuis, Alexander C. Griekspoor, Joost J. Neijssen, Carla A. Herberts, and Jacques J. Neefjes. MHC class I alleles and their exploration of the antigen-processing machinery. *Immunological Reviews*, 207(1):60–76, October 2005. URL https://doi.org/10.1111/j.0105-2896.2005.00305.x.

Mikkel Harndahl, Michael Rasmussen, Gustav Roder, Ida Dalgaard Pedersen, Mikael Sørensen, Morten Nielsen, and Søren Buus. Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. *European journal of immunology*, 42(6):1405–1416, 2012. URL https://doi.org/10.1002/eji.201141774.

Kamilla Kjærgaard Jensen, Massimo Andreatta, Paolo Marcatili, Søren Buus, Jason A. Greenbaum, Zhen Yan, Alessandro Sette, Bjoern Peters, and Morten Nielsen. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*, 154(3):394–406, 2018. URL https://doi.org/10.1111/imm.12889.

Kasper W Jørgensen, Michael Rasmussen, Søren Buus, and Morten Nielsen. NetMHCstab–predicting stability of peptide–MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology*, 141(1):18–26, 2014. URL https://doi.org/10.1111/imm.12160.

Vanessa Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters, and Morten Nielsen. NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *Journal of immunology (Baltimore, Md. : 1950)*, 199(9): 3360–3368, Nov 2017. ISSN 1550-6606. URL https://doi.org/10.4049/jimmunol.1700893.

Tamalika Kar, Utkarsh Narsaria, Srijita Basak, Debashrito Deb, Filippo Castiglione, David M Mueller, and Anurag P Srivastava. A candidate multi-epitope vaccine against SARS-CoV-2. *Scientific reports*, 10(1):1–24, 2020. URL https://doi.org/10.1038/s41598-020-67749-1.

Jan Klein and Akie Sato. The HLA system. *New England Journal of Medicine*, 343(10):702–709, 2000. URL https://doi.org/10.1056/NEJM200009073431006.

Timothy J. O'Donnell, Alex Rubinsteyn, Maria Bonsack, Angelika B. Riemer, Uri Laserson, and Jeff Hammerbacher. MHCflurry: Open-source class I MHC binding affinity prediction. *Cell Systems*, 7(1):129–132.e4, July 2018. URL https://doi.org/10.1016/j.cels.2018.05.014.

Marek Prachar, Sune Justesen, Daniel Bisgaard Steen-Jensen, Stephan Thorgrimsen, Erik Jurgons, Ole Winther, and Frederik Otzen Bagger. Identification and validation of 174 COVID-19 vaccine candidate epitopes reveals low performance of common epitope prediction tools. *Scientific Reports*, 10(1), November 2020. URL https://doi.org/10.1038/s41598-020-77466-4.

Sergio Quiñones-Parra, Emma Grant, Liyen Loh, Thi H. O. Nguyen, Kristy-Anne Campbell, Steven Y. C. Tong, Adrian Miller, Peter C. Doherty, Dhanasekaran Vijaykrishna, Jamie Rossjohn, Stephanie Gras, and Katherine Kedzierska. Preexisting CD8+ T-cell immunity to the H7N9 influenza A virus varies across ethnicities. *Proceedings of the National Academy of Sciences*, 111 (3):1049–1054, 2014. URL https://doi.org/10.1073/pnas.1322229111.

Ugur Sahin, Katalin Karikó, and Özlem Türeci. mRNA-based therapeutics — developing a new class of drugs. *Nature Reviews Drug Discovery*, 13(10):759–780, September 2014. URL https://doi.org/10.1038/nrd4278.

Tilman Schneider-Hohendorf, Dennis Görlich, Paula Savola, Tiina Kelkka, Satu Mustjoki, Catharina C. Gross, Geoffrey C. Owens, Luisa Klotz, Klaus Dornmair, Heinz Wiendl, and Nicholas Schwab. Sex bias in MHC I-associated shaping of the adaptive immune system. *Proceedings of the National Academy of Sciences*, 115(9):2168–2173, 2018. URL https://doi.org/10.1073/pnas.1716146115.

Nilabh Shastri, Sylvain Cardinaud, Susan R. Schwab, Thomas Serwold, and Jun Kunisawa. All the peptides that fit: the beginning, the middle, and the end of the MHC class I antigen-processing pathway. *Immunological Reviews*, 207(1):31–41, October 2005. URL https://doi.org/10.1111/j.0105-2896.2005.00321.x.

Johanna Vielhaben, Markus Wenzel, Wojciech Samek, and Nils Strodthoff. USMPep: universal sequence models for major histocompatibility complex binding affinity prediction. *BMC Bioinformatics*, 21(1), July 2020. URL https://doi.org/10.1186/s12859-020-03631-1.

Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343, 10 2018. ISSN 0305-1048. URL https://doi.org/10.1093/nar/gky1006.

Jatin M Vyas, Annemarthe G Van der Veen, and Hidde L Ploegh. The known unknowns of antigen processing and presentation. *Nature Reviews Immunology*, 8(8):607–618, 2008. URL https://doi.org/10.1038/nri2368.

Yiska Weisblum, Fabian Schmidt, Fengwen Zhang, Justin DaSilva, Daniel Poston, Julio CC Lorenzi, Frauke Muecksch, Magdalena Rutkowska, Hans-Heinrich Hoffmann, Eleftherios Michailidis, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife*, 9: e61312, 2020. URL https://doi.org/10.1101/2020.07.21.214759.

Marek Wieczorek, Esam T. Abualrous, Jana Sticht, Miguel Álvaro Benito, Sebastian Stolzenberg, Frank Noé, and Christian Freund. Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation. *Frontiers in Immunology*, 8: 292, 2017. ISSN 1664-3224. URL https://doi.org/10.3389/fimmu.2017.00292.

Robert J. Winchester. 5 - the major histocompatibility complex. In Robert R. Rich, Thomas A. Fleisher, William T. Shearer, Harry W. Schroeder, Anthony J. Frew, and Cornelia M. Weyand (eds.), *Clinical Immunology*, pp. 79–90. Mosby, Edinburgh, third edition, 2008. ISBN 978-0-323-04404-2. URL https://doi.org/10.1016/B978-0-323-04404-2.10005-3.

Weilong Zhao and Xinwei Sher. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. *PLOS Computational Biology*, 14(11):e1006457, November 2018. URL https://doi.org/10.1371/journal.pcbi.1006457.