

MIT Open Access Articles

Quantification of Uncertainty in Peptide-MHC Binding Prediction Improves High-Affinity Peptide Selection for Therapeutic Design

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Zeng, Haoyang and David K. Gifford. "Quantification of Uncertainty in Peptide-MHC Binding Prediction Improves High-Affinity Peptide Selection for Therapeutic Design." *Cell Systems* 9, 2 (August 2019): P159-166.e3 © 2019 Elsevier Inc

As Published: <http://dx.doi.org/10.1016/j.cels.2019.05.004>

Publisher: Elsevier BV

Persistent URL: <https://hdl.handle.net/1721.1/128919>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-NonCommercial-NoDerivs License





Published in final edited form as:

Cell Syst. 2019 August 28; 9(2): 159–166.e3. doi:10.1016/j.cels.2019.05.004.

Quantification of uncertainty in peptide-MHC binding prediction improves high-affinity peptide selection for therapeutic design

Haoyang Zeng^{1,2}, David K. Gifford^{1,2,*}

¹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

²Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

Summary

The computational identification of peptides that can bind the major histocompatibility complex (MHC) with high affinity is an essential step in developing personal immunotherapies and vaccines. We introduce PUFFIN, a deep residual network-based computational approach that quantifies uncertainty in peptide-MHC affinity prediction that arises from observational noise and the lack of relevant training examples. With PUFFIN's uncertainty metrics, we define binding likelihood, the probability a peptide binds to a given MHC allele at a specified affinity threshold. Compared to affinity point estimates, we find that binding likelihood correlates better with the observed affinity and reduces false positives in high-affinity peptide design. When applied to examine an existing peptide vaccine, PUFFIN identifies an alternative vaccine formulation with higher binding likelihood. PUFFIN is freely available for download at <http://github.com/gifford-lab/PUFFIN>.

Graphical Abstract

Machine learning models that predict the binding affinity of a peptide-MHC pair are essential in peptide-based therapeutic design, but state-of-the-art methods provide point estimates of affinity that do not consider measurement noise and model uncertainty. We introduce PUFFIN, a method that quantifies the prediction uncertainty and prioritizes peptides with *binding likelihood* to achieve improved accuracy in high-affinity peptide selection for therapeutic design.

*Corresponding Author and Lead Contact: D.K.G. (gifford@mit.edu).

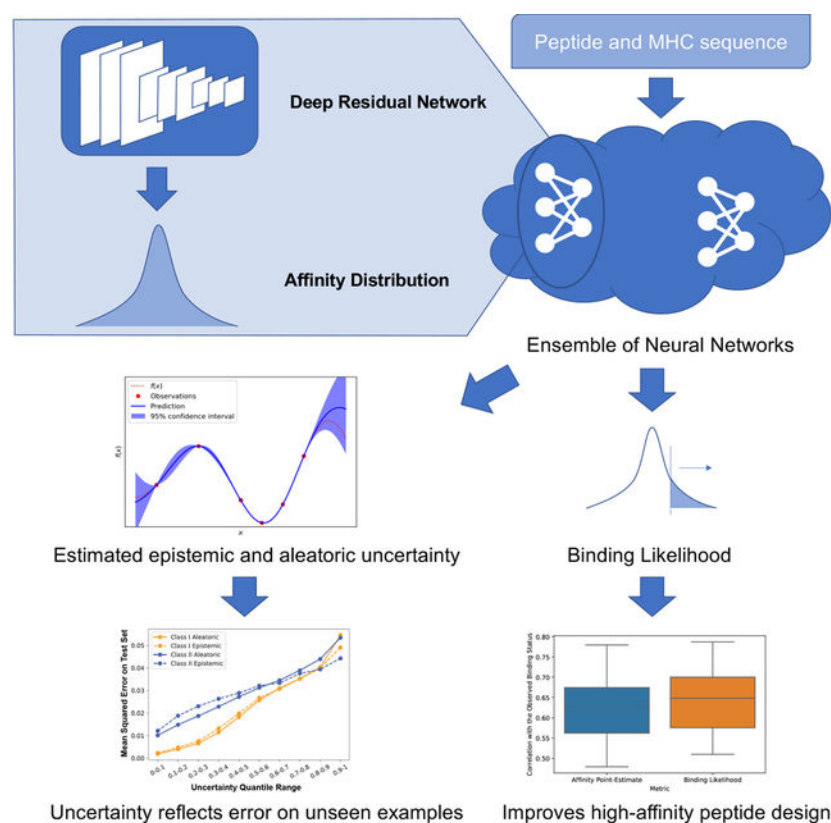
Author Contributions

H.Z. designed the study, with input from D.K.G. H.Z. and D.K.G. developed the method and analyzed the results. D.K.G. supervised the study. H.Z. and D.K.G. wrote the paper.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Competing Interests

D.K.G. is a founder of Think Therapeutics, a company that uses machine learning for therapeutic design.



Introduction

The major histocompatibility complex (MHC) is a set of cell surface proteins that are crucial for the extra-cellular display of peptides for surveillance by the immune system (Castellino et al., 1997; Janeway Jr et al., 2001). Peptides displayed by MHC molecules are either synthesized in the cell (class I MHC) or internalized from the extracellular medium (class II MHC). T cells routinely surveil the peptides presented on cell surfaces and trigger an immune response upon recognition of non-self peptides that arise from foreign antigens, cell infection, or mutated self-proteins. The selective binding of peptides to MHC molecules plays an essential role in individual-specific peptide presentation to the immune system. Understanding what peptides will be displayed given a disease state is crucial information for developing peptide-based vaccines and therapeutics (Kreiter et al., 2015; Ott et al., 2017; 2017; Verdegaaal et al., 2016).

MHC molecules are encoded by highly polymorphic gene families (Jin and Wang, 2003; Robinson et al., 2015; Williams, 2001). Each MHC genotype has its own specificity for peptide presentation. Moreover, the length of the peptides that bind to an MHC molecule varies. Class I MHC presented peptides are typically 8 to 11 amino acids (Lundegaard et al., 2008), while class II MHC presented peptides are typically 13 to 25 amino acids (Chicz et al., 1992). In class II MHC molecules a core 9-amino-acid sequence interacts with the class II MHC groove while the peptide regions outside of the binding have a secondary influence on binding affinity (Arnold et al., 2002; Holland et al., 2013).

Machine learning models for MHC-peptide affinity are essential for certain therapeutic applications including predicting cancer-specific peptides that will be presented by the MHC on tumor cells and engineering peptide vaccines (Kreiter et al., 2015; Ott et al., 2017; 2017). The task of these machine learning models is to input a peptide and an MHC sequence and output the binding affinity of the peptide to the MHC molecule. Two widely used methods, NetMHCpan (Nielsen and Andreatta, 2016) (for class I MHC) and NetMHCIIpan (Jensen et al., 2018) (for class II MHC), use a one-layer fully-connected neural network to iteratively identify a peptide's 9-mer binding core and predict its binding affinity (Nielsen and Andreatta, 2017). Other recent methods for class I MHC binding use shallow convolutional or recurrent neural networks (Bhattacharya et al., 2017; Han and Kim, 2017; O'Donnell et al., 2018; Vang and Xie, 2017).

Accurate machine learning models for predicting MHC-peptide affinity depend upon sufficient training data to allow them to generalize to unseen inputs. At present model generalization is constrained by the limited size of current training datasets. The largest database of MHC-peptide binding, Immune Epitope Database and Analysis Resource (IEDB) (Vita et al., 2018), has affinity data for over 80 human and mouse alleles for class II MHC. However, only six MHC alleles have more than 5000 peptide examples, and the most abundant allele has only 10,000 examples, a tiny fraction of all possible peptides of similar lengths. Moreover, IEDB entries harbor strong measurement noise that arises from batch effects and protocol differences since they are curated from published reports from disparate laboratories. For a given input, the sufficiency of relevant training data and the level of measurement noise can be quantified by proper uncertainty metrics associated with each computational prediction (Kendall and Gal, 2017). Despite the promising results of existing machine learning methods, they do not provide an uncertainty metric for their point estimate of MHC-peptide affinity.

Here we introduce PUFFIN (Prediction of Uncertainty in MHC-peptide affinity using residual Networks), a method for predicting MHC-peptide binding that outputs both the *expected affinity* of an input MHC-peptide pair as well as the *uncertainty* of the model about its prediction. PUFFIN has the dual advantage of providing more accurate estimates of both class I and class II MHC-peptide binding than previous methods as well as uncertainty metrics with each prediction. Our prediction uncertainty metrics characterize the uncertainty that results from inherent observational noise (aleatoric uncertainty) and bounded knowledge about model selection (epistemic uncertainty). PUFFIN's uncertainty metrics allows for the principled comparison of alternative peptide vaccine formulations, and permits peptides to be selected based upon *binding likelihood*, the probability that a peptide binds to a given MHC molecule at a specified affinity threshold. Compared to the conventional approach that relies on point estimates of affinity, our proposed approach reduces false positives in high-affinity peptide design.

Results

Modeling both epistemic and aleatoric uncertainty

PUFFIN provides distinct estimates of model uncertainty (epistemic) and inherent observational uncertainty (aleatoric). Epistemic uncertainty refers to the uncertainty that

arises from the lack of model knowledge and can be reduced with additional training data (Kendall and Gal, 2017). Given a fixed-size dataset, alternative models might exist to explain the observed examples well. Epistemic uncertainty describes the uncertainty that one has about the true model that generates the data, and is higher for inputs where relevant training observations are lacking. Aleatoric uncertainty refers to the uncertainty that results from the inherent noise in observations and thus cannot be reduced with more training examples (Kendall and Gal, 2017). Aleatoric uncertainty can be further categorized into homoscedastic and heteroscedastic uncertainty. Heteroscedastic uncertainty is observational noise that depends on the input value, while homoscedastic uncertainty remains constant for all inputs and thus can be considered as a special and simpler version of heteroscedastic uncertainty.

We model both the epistemic and heteroscedastic aleatoric uncertainty in MHC-peptide binding affinity using a unified framework. We model the affinity of an MHC-peptide pair as a random sample from a probability distribution (beta distribution for class I, normal distribution for class II, chosen by empirical performance in cross-validation) where the distribution's parameters are predicted from the input MHC and peptide sequence using a deep residual convolutional network (He et al., 2016) (Figure 1, Methods). The dispersion of the probability distribution characterizes the aleatoric uncertainty specific to the corresponding input. Epistemic uncertainty reflects the existence of alternative models that explain the observed data. The most established way to obtain epistemic uncertainty is through the posterior distribution of model parameters in a Bayesian learning framework (Kendall and Gal, 2017). Certain model selection procedures such as Stability Selection (Meinshausen and Bühlmann, 2010) also implicitly choose the model with the highest posterior probability, but the optimal model chosen from such procedures can't provide uncertainty metrics associated with a prediction. Ensemble models have been established as a Bayesian approximation and are more scalable compared to Bayesian neural networks (Lakshminarayanan et al., 2017). Thus, we characterize epistemic uncertainty by the predictive variance across an ensemble of neural network models trained on different training-validation splits and with different random initializations (Methods).

PUFFIN's affinity point estimates demonstrates state-of-the-art accuracy

We first show that PUFFIN's mean estimates of affinity outperform previous state-of-the-art models for MHC-peptide affinity prediction. We average the predicted means from all PUFFIN ensemble members to produce PUFFIN's mean estimate prediction and compare it with existing methods.

For class II MHC, we trained and evaluated PUFFIN and NetMHCIIpan on binding affinity data as per Jensen et al. (Jensen et al., 2018). Jensen et al. curated class II MHC-peptide affinity data from IEDB and split it into five cross-validation folds such that different folds do not share 9-mer peptide sequences (Jensen et al., 2018; Nielsen et al., 2007). We made predictions for each fold using a model trained the other four folds so that the performance was evaluated on held out examples. The re-training of NetMHCIIpan model on this dataset is enabled by a standalone training platform published by Nielsen et al. (Nielsen and Andreatta, 2017). We evaluated the prediction performance using auROC, F1 score, mean-

squared-error (MSE), R^2 , Spearman correlation, and Point-Biserial correlation. For auROC, F1 score and Point-Biserial correlation, positive examples were defined as the ones with a binding affinity stronger than 500 nM as used in the literature (Bhattacharya et al., 2017; Jensen et al., 2018; Nielsen and Andreatta, 2016). We found that PUFFIN's mean estimate outperforms NetMHCIIpan in all metrics considered when evaluated on all MHC-peptide pairs (Table 1). Combining the predictions from PUFFIN and NetMHCIIpan yields further performance improvement, suggesting complimentary features might be captured by the two approaches. Furthermore, when evaluated on each MHC allele separately, PUFFIN has a lower mean-squared-error than NetMHCIIpan for 44 of the 55 MHC alleles considered (Figure 2A).

For class I MHC, the Nielsen et al (Nielsen and Andreatta, 2017) training platform is not able to fully reproduce the same training procedure as published for NetMHCpan (Nielsen and Andreatta, 2016). Bhattacharya et al. (Bhattacharya et al., 2017) provided a benchmark dataset on which several recent computational methods for class I MHC-peptide binding were evaluated and their performance was reported (Table 2). Trained and tested on the same benchmark, PUFFIN outperforms all the competing methods including NetMHCpan (Nielsen and Andreatta, 2016), MHCflurry (O'Donnell et al., 2018), and MHCNugget (Bhattacharya et al., 2017) in auROC and Kendall's tau and shows competitive performance in F1 score (Table 2).

Uncertainty accurately reflects the predictive error in affinity prediction

We next show that uncertainty estimates from PUFFIN provide a way to gauge the predictive error on unseen examples. Uncertainty characterizes the lack of confidence in a prediction caused by either the lack of model selection and training data (epistemic uncertainty) or observation noise (aleatoric uncertainty) near the queried data point. For reliable uncertainty estimates, the level of confidence should match the predictive accuracy on a held-out dataset and indicate how much one can trust a prediction (Lakshminarayanan et al., 2017).

To examine the quality of PUFFIN's uncertainty estimation, we predicted the affinity of all the examples in IEDB in the same cross-validation manner as described above. For each MHC-peptide pair, the mean and variance of the affinity distribution predicted by each of the networks in PUFFIN's ensemble were calculated. Across all networks in PUFFIN's ensemble, the average of the affinity variances was used to quantify the aleatoric uncertainty, and the variance of the affinity means was used to quantify the epistemic uncertainty. For each type of uncertainty, we binned the held-out test examples according to their uncertainty quantiles and calculated the mean-squared-error in each bin. We observed that both PUFFIN's epistemic and aleatoric uncertainty highly correlate with the prediction error, and predictions made with lower uncertainty are more accurate (Figure 2B). This faithful stratification of predictive performance on held-out observations demonstrates that PUFFIN's uncertainty estimations reflect its predictive confidence and provides useful guidance on how to utilize its computational predictions.

Epistemic uncertainty identifies sequences foreign to the model

We next show that PUFFIN identifies sequences foreign to the model by labeling them with high epistemic uncertainty. Here we assume that reliable epistemic uncertainty should increase for examples that are distant from a model's training examples to reflect the absence of relevant training data.

We first created a systematic survey of PUFFIN's predicted epistemic uncertainty by characterizing examples at varying edit distances from a model's training examples. Given a training set and a test set, we randomly sampled from the training set 10,000 MHC-peptide pairs where the peptides were at least 10 amino acids long. These peptides were used as "seed" sequences with known affinity to their respective MHC. We then created three sets of derived peptide sequences that are respectively 1, 5 or 10 amino acids different from the seeds. Specifically, for each seed, 10 sequences with the desired number of mutations (1 or 5 or 10) were randomly created. We assume that the designed distance to a seed approximates the distance to the whole training set when the sequence space is large. We focused on class II MHC as its longer input peptide sequences result in a larger sequence space than class I MHC.

We applied PUFFIN to predict the binding affinity of the seed sequences as well as the three sets of derived sequences using the five-fold cross-validation data from Jensen et al. Each cross-validation fold yields a distinctive split of training and test set, on which the seeds and three sets of derived sequences were generated as described above and evaluated on a PUFFIN model trained on the corresponding training set. We calculated the median epistemic and aleatoric uncertainty for all sequences in each set. To account for differences among the seeds, we also adopted an alternative metric: for mutation sequences derived from the same seed, we computed the median uncertainty as a representative and reported the median of representative uncertainty across all seeds. Under both metrics, we observed that sequences with increasing distance from seeds result in as much as 35% more epistemic uncertainty (Figure 2C), showing that PUFFIN's epistemic uncertainty characterizes the lack of relevant training examples in the neighborhood of an input. Meanwhile, we found that aleatoric uncertainty remains largely unchanged as expected. These results support our hypothesis that the epistemic and aleatoric uncertainty estimates from PUFFIN respectively correspond to the uncertainty that results from a lack of model knowledge and from observational noise.

Uncertainty estimation improves precision in high-affinity peptide design

In the design of high-affinity peptides for therapeutic purposes, the number of distinct peptides administered is constrained (Kreiter et al., 2015; Ott et al., 2017; 2017) and reducing the false selection of peptides is essential to achieve high efficacy with a minimal dosage. False positives arise from erroneous model predictions of high peptide affinity. False predictions are not possible to detect when models do not provide uncertainty estimates. Unlike the contemporary models, PUFFIN computes the *binding likelihood* of a peptide, defined as the probability the peptide binds to a given MHC allele at a specified affinity threshold (Methods). We found that prioritizing peptides based on binding likelihood leads to improved precision in high-affinity peptide design.

We first hypothesized that binding likelihood at an affinity threshold of 500 nM enables a more accurate prediction of observed binding status defined by a 500 nM affinity threshold. For each allele in the held-out dataset, we scored a peptide by both the predicted affinity and the likelihood that the observed affinity is at least as strong as 500 nM under PUFFIN's probabilistic model. Point-Biserial correlations between observed binding and our two predicted metrics, binding affinity and binding likelihood, were calculated respectively for each MHC allele. For both class I and class II MHC, we found that observed binding correlates better with binding likelihood than the predicted mean affinity for all MHC alleles with more than 2000 examples (Figure 3A, 32/32 for class I MHC and 28/28 for class II MHC) and for the majority of all MHC alleles examined (94/113 for class I MHC, and 51/55 for class II MHC; Methods). Across all MHC alleles, binding likelihood shows a statistically significant improvement over predicted mean affinity in correlation (Wilcoxon one-sided signed rank test; $p=1.2e-09$ for class I MHC; $p=1.2e-08$ for class II MHC). When we evaluated PUFFIN on the Bhattacharya et al. (Bhattacharya et al., 2017) benchmark data as described above, we also observed that PUFFIN's binding likelihood further improves the auROC and Point-Biserial correlation (Table 1, 2). As expected, we didn't observe a stronger Spearman correlation or Kendall's Tau which are metrics that quantify the fit to affinity values rather than the discrimination of binding status with respect to an affinity cutoff.

We next hypothesized that binding likelihood would lead to more accurate identification of MHC-binding peptides. This is because PUFFIN makes it possible to evaluate the reliability of computationally predicted affinities and select peptides that are predicted to bind with high confidence. For each MHC allele, we identified the peptides from a held-out candidate set with a PUFFIN predicted binding likelihood above 95%. For these peptides, failing to bind to the target MHC molecule is unlikely under PUFFIN's probabilistic model. Only MHC alleles with over 100 training peptides were considered in view of the noise present in affinity measurements. We also identified the same number of peptides with the highest predicted mean affinity for each MHC allele to compare with a conventional strategy based solely on high-affinity peptides. For class I MHC, we observed that for 80% of the 15 MHC alleles considered, the percentage of true binders (stronger than 500 nM) among the peptides identified by PUFFIN is higher than (for 46.7% of the alleles) or equal to (for 33.3% of the alleles) that for the peptides with the highest affinity prediction. For class II MHC, the peptides predicted to bind by PUFFIN with high-confidence are more likely (for 60.8% of the alleles) or equal likely (for 17.3% of the alleles) to bind than peptides predicted with the highest affinity for a total of 78.1% of the 23 alleles considered (Figure 3B).

Published peptide vaccine formulations can be improved by uncertainty metrics

We next applied PUFFIN's binding likelihood estimates to examine the peptide formulation of neoantigen vaccines for melanoma. Ott et al. (Ott et al., 2017) designed personal neoantigen vaccines to induce tumor-specific T-cell responses to melanoma for six patients. Somatic mutations were identified from whole exon sequencing data from tumor and germline cells. NetMHCpan's computational predictions were used to rank mutation-spanning peptides by the binding affinity to patient-specific MHC class I molecules. We applied PUFFIN to examine the binding likelihood (target affinity of 500 nM) of each peptide in the vaccine to the patient-specific MHC class I allele. We observed that the

median binding likelihood of the peptides in the published vaccines range from 51.7% to 73.6%, with certain peptides having a binding likelihood below 20% for five of the six vaccines (Figure 3C). The low binding likelihoods we observed suggested room for improvement in the prioritization of MHC-binding peptides. For each patient, we examined the binding likelihood between all 9-mer peptides that span the somatic mutations and the MHC alleles of the same patient. The median binding likelihoods of mutation-spanning peptides for each patient are low, ranging from 9×10^{-7} to 0.0022 (Figure 3D). We found that with the same peptide count an alternative set of peptides exists with significantly higher binding likelihoods than in the published vaccines (Figure 3C). The proposed set of peptides could potentially lead to an improved rate of T cell response compared to the existing vaccine candidates that were selected by affinity point estimates predicted by conventional computational methods. This result suggests that uncertainty will be a useful metric for peptide vaccine formulation, but further testing in the context of additional constraints in a clinical setting will be necessary to confirm the utility of uncertainty for vaccine formulation.

IEDB power by allele and MHC class

We next applied PUFFIN to characterize the power of the IEDB datasets to predict peptide-MHC binding affinity. As a database curated from publications in which different experimental protocols and conditions were employed, the IEDB datasets are inherently noisy.

Moreover, the number of available examples for different MHC alleles is highly skewed towards a few common alleles, leading to variability in the predictive power of computational models across MHC alleles. Thus, we used PUFFIN to examine how the epistemic and aleatoric uncertainty changes across MHC alleles.

We first show that binding affinity data for class II MHC is more heterogeneous than class I (Figure 4A). For class I MHC, we found the correlation between the median aleatoric uncertainty and the dataset size is not statistically significant (Pearson $r=-0.11$, $p=0.26$). A stronger and statistically significant correlation was observed for class II (Pearson $r=0.3$, $p=0.027$), indicating that examples for the alleles with more data tend to harbor higher inherent noise. We highlight the class II MHC alleles with the top median aleatoric uncertainty in Table S1. We note that certain alleles, such as HLA-DQA10102-DQB10501, H-2-IA δ , and HLA-DQA10201-DQB10301, have high aleatoric uncertainty with small datasets.

We found that a larger training set size improves prediction confidence for class I MHC but not for class II MHC (Figure 4B). For class I MHC, we found a strong negative correlation between the median epistemic uncertainty and the dataset size (Pearson $r=-0.38$, $p=3.4e-05$), suggesting that the prediction confidence on the held-out test set increases with the size of the training set. In contrast, no correlation was observed for class II (Pearson $r=-0.065$, $p=0.64$). This difference in the correlation to dataset size could result from the fact that peptides that bind to class II MHC span a much greater sequence space due to a larger range of peptide length. Thus, significantly more training data are required to sufficiently fill

the sequence space and thus lead to a decrease in epistemic uncertainty for predictions made for held-out examples.

Discussion

Predicting the binding of peptides to MHC molecules is a central task in characterizing the antigens T cells can surveil and the design of personal peptide-based therapeutics. The performance of computational models are bounded by the limited size of available training data sets for most MHC alleles. We have found that model uncertainty metrics are a useful adjunct to predicted mean affinity, and can be used to compute a binding likelihood metric for peptide selection in vaccine formulation.

Unlike existing methods such as NetMHCpan and MHCflurry, PUFFIN provides uncertainty estimates for MHC-peptide affinity prediction that provides uncertainty estimates. We show that PUFFIN's uncertainty estimates are able to reflect the predictive error on unseen examples. PUFFIN's epistemic uncertainty output characterizes model uncertainty and identifies the sequences far away from the training examples, while PUFFIN's aleatoric uncertainty output characterizes the inherent noise in the measurement. Compared to existing point-estimate methods, PUFFIN enables a probabilistic characterization of the binding affinity that conveys the confidence in the prediction. This probabilistic framework allows us to define binding likelihood, the probability that a peptide binds to a MHC molecule at a given affinity threshold, as a metric that facilitates precise prioritization of high-affinity peptides. This prioritization is a central task in peptide-based therapeutic design. Binding likelihood analysis of a published peptide vaccine suggests room for improvement in selecting mutation-spanning peptides with higher MHC-binding affinity and thus greater chances of being T cell epitopes.

Our work demonstrates the importance of incorporating uncertainty estimation in the design of computational models for MHC-peptide binding. A metric of uncertainty not only leads to more accurate models for important therapeutic applications, but also helps reveal the inherent noise in measurement and the limitations of computational models. Reliable uncertainty estimation can also enable therapeutic applications that are not previously possible, such as the computational optimization of peptides for binding affinity and specificity. In such applications, uncertainty estimates can be employed by Bayesian Optimization to efficiently explore sequence space for an optimum candidate.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, David Gifford (gifford@mit.edu).

METHOD DETAILS

MHC-peptide binding affinity dataset—We used MHC-peptide binding affinity dataset deposited in The Immune Epitope Database (IEDB). The affinities of peptides and MHC

molecules are represented as IC50 values, the half-maximal inhibitory concentration in nano-molar (nM) units determined by immunofluorescence assays.

Using class I MHC-peptide affinity dataset from IEDB, Bhattacharya et al.(Bhattacharya et al., 2017) constructed a benchmark in which no peptide in the test set has identical length and greater than 80% sequence identity to any peptide in the training set. 51 class I MHC alleles are covered in this dataset. We used the same training and test set in this benchmark to compare PUFFIN with the existing computational methods. For the rest of the analyses that involve class I MHC-peptide binding, we used the IEDB-based dataset collected by Nielsen et al.(Nielsen and Andreatta, 2016) in which five cross-validation folds were created to ensure no peptide shares a 9-mer sequence with any peptide in a different fold. By training a model on four folds and test on the remaining fold, this setup allows us to analyze the performance on all pairs of MHC and peptide available in the dataset in a cross-validation manner. For class II MHC-peptide binding, we used the IEDB-based dataset that Jensen et al.(Jensen et al., 2018) collected in which five cross-validation folds were created in the same way as in Nielsen et al. For the latter two datasets, only MHC alleles (114 for class I MHC and 55 for class II MHC) with more than 100 examples were included to ensure the quality of training. In all three datasets, the IC50 values have been normalized by $1 - \log(\text{IC50})/\log(50000)$ to be between 0 and 1.

Feature representation—PUFFIN takes as input a MHC-peptide pair and predicts a probabilistic distribution of peptide-MHC binding affinity. As was used in past literature, each MHC allele was represented by a pseudo-sequence consisting of 34 amino acid residues in contact with the peptide(Jensen et al., 2018; Karosiene et al., 2013; Nielsen and Andreatta, 2016). The contact residues were defined as the polymorphic residues that are within 4.0 Å of the peptide in the structure of one or more of major MHC alleles. All peptides sequences were padded on the right end to the same length, 30 for class I and 40 for class II, using a place-holder amino acid to work with the maximum length of peptides in IEDB.

Each amino acid was encoded as a feature vector of 40 values concatenated from two representations: a one-hot encoding vector of 20 values to denote the 20 amino acids of interest; and the row (of 20 values) of BLOSUM50 (Nielsen et al., 2003) matrix that corresponds to the amino acid to characterize the evolutionary similarities between amino acids. A one-hot encoding represents a categorical variable by a vector of which the position that corresponds to the observed category is one and all other positions are zero. The BLOSUM50 matrix has been used in previous peptide-MHC binding prediction methods to represent amino acids(Jensen et al., 2018; Nielsen and Andreatta, 2016). We used the BLOSUM50 matrix in third-bit units from NCBI (https://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/data/BLOSUM50). For the place-holder amino acids, a dummy “one-hot” encoding vector full of zero and a dummy “BLOSUM50 row” full of the lowest substitution score were used. Thus, each MHC allele was represented as a 40×34 matrix and was reshaped to a one-dimensional feature vector of 1360 values. Each peptide was represented as a matrix of size 40×30 for class I MHC and 40×40 for class II MHC. Given a MHC-peptide pair, we embedded the MHC feature vector into the first dimension of the peptide feature matrix to form a final input matrix of size 1400 × 30 for class I MHC and

1400×40 for class II MHC. We also encoded the difference between the peptide length L and the expected length \bar{L} (9 for class I MHC and 15 for class II MHC) using a sigmoid function $\hat{L} = \frac{1}{1 + \exp((L - \bar{L})/2)}$.

Network structure and training—PUFFIN is an ensemble of deep neural networks implemented in Pytorch. For a given pair of MHC and peptide, the observed affinity is modeled as a sample from a probability distribution (beta distribution for class I, normal distribution for class II, chosen by empirical performance in cross-validation). Each network in the ensemble predicts the parameters of the affinity distribution (# and \$ for beta distribution, mean and variance for normal distribution). The mean and variance of the distribution were used to quantify uncertainty. For a given MCH-peptide pair, the ensemble-wide average of the affinity variance characterizes the aleatoric uncertainty, and the ensemble-wide variance of the affinity mean characterizes the epistemic uncertainty. To compare with existing methods that provide a point estimate of affinity, the ensemble-wide average of the affinity mean was used. For a given affinity threshold, the binding likelihood was calculated as the probability that the observed affinity is beyond the threshold using a beta (for class I MHC) or normal (for class II MHC) distribution with parameters averaged across networks in the ensemble.

Each network in the ensemble is a deep residual convolutional neural network(He et al., 2016) followed by two fully-connected layers. The residual network learns a high-level representation of the MHC-peptide pair, which is then concatenated with the sigmoid-transformed peptide length \hat{L} and $1-\hat{L}$ for the final prediction of the affinity distribution using the two fully-connected layers.

The residual network consists of an initial convolutional layer and five convolutional residual blocks. Each residual block has two convolutional layers that learn the residuals between the output and the input. Followed by a batch-normalization layer, every convolutional layer in the residual network has 256 convolutional kernels with a stride of 1. ReLU activation is used as non-linearity across the network. Training of the network was performed using the adaptive stochastic gradient descent method Adam(Kingma and Ba, 2014). For a given pair of training and test set, we randomly held out 1/8 of the training set as a validation set. All hyper-parameters, including the number of layers, the number of convolutional kernels, and the parameters of the optimizer, were chosen based on the loss on the validation set. The final training was performed for 50 epochs with early stopping if no improvement on validation loss was observed for 10 epochs. The model weights from the epoch with the lowest validation loss were selected. We generated 10 such random splits of training and validation set. For each split, two models were trained with different random weight initializations. The resulting 20 models form the final ensemble of PUFFIN to make predictions on the test set, which was held out for all networks in the ensemble. With the five-fold cross-validation datasets from Nielsen et al. and Jensen et al., the affinity of each MHC-peptide pair in a fold was predicted by a PUFFIN model trained on the other four folds. This enables us to compare the predictive performance on all examples in IEDB while ensuring that all performance was fairly evaluated.

Comparison with NetMHCIIpan for class II MHC-peptide binding—The published NetMHCIIpan3.2(Jensen et al., 2018) model is an ensemble of models trained on different subsets of the dataset from Jensen et al. To ensure a fair comparison with PUFFIN, we re-trained NetMHCIIpan3.2 using the NNAlign-2.0 platform(Nielsen and Andreatta, 2017) developed by the same lab. The same network structures and hyper-parameters as described in the NetMHCIIpan3.2 publication were used. For each of the five training-test fold splits in the dataset constructed by Jensen et al., we trained PUFFIN and NetMHCIIpan3.2 on the training set and predicted on the corresponding test set. The cross-validation predictions for all examples in the dataset were combined and evaluated against the observed affinities.

Correlation analysis with observed affinity—113 of the 114 class I MHC alleles in the data from Nielsen et al. were included in the comparison of binding likelihood and predicted mean affinity's correlations with the observed affinity because all examples of HLA-B27:03 have the same observed affinity and thus the corresponding Pearson correlation is ill-defined.

Analysis of published neoantigen vaccine—Ott et al.(Ott et al., 2017) released the sequence and the target MHC of all the peptides designed in their vaccines as well as the somatic mutations of each of the six patients. For peptides selected for high predicted affinity (stronger than 500 nM), we evaluated the binding likelihood to their targeted MHC. To construct an alternative set of peptides with improved binding likelihood, we identified all the 9-mer peptides that span any missense somatic mutation. Mutations that reside in UTRs, non-coding regions, and introns, as well as the ones that lead to frameshift were excluded as their effects on introducing novel peptides cannot be properly evaluated without the availability of RNA-seq data. For each patient, the number of mutation-spanning peptides evaluated are 697, 7943, 3141, 8438, 2869 and 15021 respectively. The binding likelihoods between the mutation-spanning peptides and the MHCs of the same patient were predicted with PUFFIN.

QUANTIFICATION AND STATISTICAL ANALYSIS

The Point-Biserial correlations between the observed binding status and the computational predictions, either binding likelihood or predicted mean affinity, were calculated by Python package *scipy.stats.pointbiserialr*. The Wilcoxon one-sided signed rank tests to compare the Point-Biserial correlations calculated for binding likelihood and predicted mean affinity were performed using R package *stats.wilcox.test*. The number of samples in such tests is 113 for class I MHC and 55 for class II MHC. The Pearson correlations between uncertainty metrics and the training set size for a MHC allele in IEDB was calculated using Python package *scipy.stats.pearsonr*. The number of samples in such tests is 114 for class I MHC and 55 for class II MHC. For all auROC and auPRC calculation, the Python package *scikit-learn* is used (*sklearn.metrics.roc_auc_score*, *sklearn.metrics.average_precision_score*). Throughout the paper, statistical significance is defined as $p < 0.05$.

DATA AND SOFTWARE AVAILABILITY

The code for PUFFIN is available at <http://github.com/gifford-lab/PUFFIN>. We obtained IEDB MHC-peptide binding affinity data from <http://www.cbs.dtu.dk/services/>

NetMHCpan-3.0/ and <http://www.cbs.dtu.dk/services/NetMHCIIpan-3.2/> for class I and class II MHC respectively. We obtained data in the class I MHC-peptide binding affinity benchmark from personal correspondence with Bhattacharya et al. and we have deposited this dataset in Mendeley Data at <https://doi.org/10.17632/jwhmrxdx268.1>

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

We would like to thank Rohit Bhattacharya for providing the benchmark dataset for class I MHC-peptide binding. We appreciate the instructions from Kamilla Kjærgaard Jensen and Morten Nielsen on training NetMHCIIpan. This work was supported by National Institute of Health grant R01CA218094.

References

- Arnold PY, La Gruta NL, Miller T, Vignali KM, Adams PS, Woodland DL, and Vignali DAA (2002). The majority of immunogenic epitopes generate CD4+ T cells that are dependent on MHC class II-bound peptide-flanking residues. *J. Immunol* 169, 739–749. [PubMed: 12097376]
- Bhattacharya R, Sivakumar A, Tokheim C, Guthrie VB, Anagnostou V, Velculescu VE, and Karchin R (2017). Evaluation of machine learning methods to predict peptide binding to MHC Class I proteins. *BioRxiv* 154757.
- Castellino F, Zhong G, and Germain RN (1997). Antigen presentation by MHC class II molecules: invariant chain function, protein trafficking, and the molecular basis of diverse determinant capture. *Hum. Immunol* 54, 159–169. [PubMed: 9297534]
- Chicz RM, Urban RG, Lane WS, Gorga JC, Stern LJ, Vignali DAA, and Strominger JL (1992). Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* 358, 764–768. [PubMed: 1380674]
- Han Y, and Kim D (2017). Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. *BMC Bioinformatics* 18, 585. [PubMed: 29281985]
- He K, Zhang X, Ren S, and Sun J (2016). Deep Residual Learning for Image Recognition *Proc. IEEE Conf. Comput. Vis. Pattern Recognit* 770–778.
- Holland CJ, Cole DK, and Godkin A (2013). Re-Directing CD4+ T Cell Responses with the Flanking Residues of MHC Class II-Bound Peptides: The Core is Not Enough. *Front. Immunol* 4, 172. [PubMed: 23847615]
- Janeway CA Jr, Travers P, Walport M, and Shlomchik MJ (2001). The major histocompatibility complex and its functions In *Immunobiology: The Immune System in Health and Disease* 5th Edition, (Garland Science), p.
- Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, Sette A, Peters B, and Nielsen M (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* 154, 394–406. [PubMed: 29315598]
- Jin P, and Wang E (2003). Polymorphism in clinical immunology-From HLA typing to immunogenetic profiling. *J. Transl. Med* 1, 8. [PubMed: 14624696]
- Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, and Nielsen M (2013). NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* 65, 711–724. [PubMed: 23900783]
- Kendall A, and Gal Y (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, pp. 5574–5584.
- Kingma DP, and Ba J (2014). Adam: A Method for Stochastic Optimization. *ArXiv Prepr ArXiv1412.6980*.

- Kreiter S, Vormehr M, van de Roemer N, Diken M, Löwer M, Diekmann J, Boegel S, Schrörs B, Vascotto F, Castle JC, et al. (2015). Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature* 520, 692–696. [PubMed: 25901682]
- Lakshminarayanan B, Pritzel A, and Blundell C (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles In *Advances in Neural Information Processing Systems* 30, (Curran Associates, Inc.), pp. 6402–6413.
- Lundegaard C, Lund O, and Nielsen M (2008). Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics* 24, 1397–1398. [PubMed: 18413329]
- Meinshausen N, and Bühlmann P (2010). Stability selection. *J. R. Stat. Soc. Ser. B (Statistical Methodol)* 72, 417–473.
- Nielsen M, and Andreatta M (2016). NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med* 8, 33. [PubMed: 27029192]
- Nielsen M, and Andreatta M (2017). NNAlign: a platform to construct and evaluate artificial neural network models of receptor–ligand interactions. *Nucleic Acids Res* 45, W344–W349. [PubMed: 28407117]
- Nielsen M, Lundegaard C, Wornig P, Lauemøller SL, Lamberth K, Buus S, Brunak S, and Lund O (2003). Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* 12, 1007–1017. [PubMed: 12717023]
- Nielsen M, Lundegaard C, and Lund O (2007). Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* 8, 238. [PubMed: 17608956]
- O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, and Hammerbacher J (2018). MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Syst* 0.
- Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, Zhang W, Luoma A, Giobbie-Hurder A, Peter L, et al. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 547, 217–221. [PubMed: 28678778]
- Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, and Marsh SGE (2015). The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* 43, 423–431.
- Vang YS, and Xie X (2017). HLA class I binding prediction via convolutional neural networks. *Bioinformatics* 33, 2658–2665. [PubMed: 28444127]
- Verdegaal EME, de Miranda NFCC, Visser M, Harryvan T, van Buuren MM, Andersen RS, Hadrup SR, van der Minne CE, Schotte R, Spits H, et al. (2016). Neoantigen landscape dynamics during human melanoma–T cell interactions. *Nature* 536, 91–95. [PubMed: 27350335]
- Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, and Peters B (2018). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res*
- Williams TM (2001). Human Leukocyte Antigen Gene Polymorphism and the Histocompatibility Laboratory. *J. Mol. Diagnostics* 3, 98–104.
- (2017). The problem with neoantigen prediction. *Nat. Biotechnol* 35, 97–97. [PubMed: 28178261]

Highlights

- Quantifies uncertainty in peptide-MHC affinity prediction
- Predicted uncertainty correlates with the observed error on held-out examples
- Uses a binding likelihood metric that improves upon point affinity predictions
- Improves high-affinity peptide selection for therapeutic design

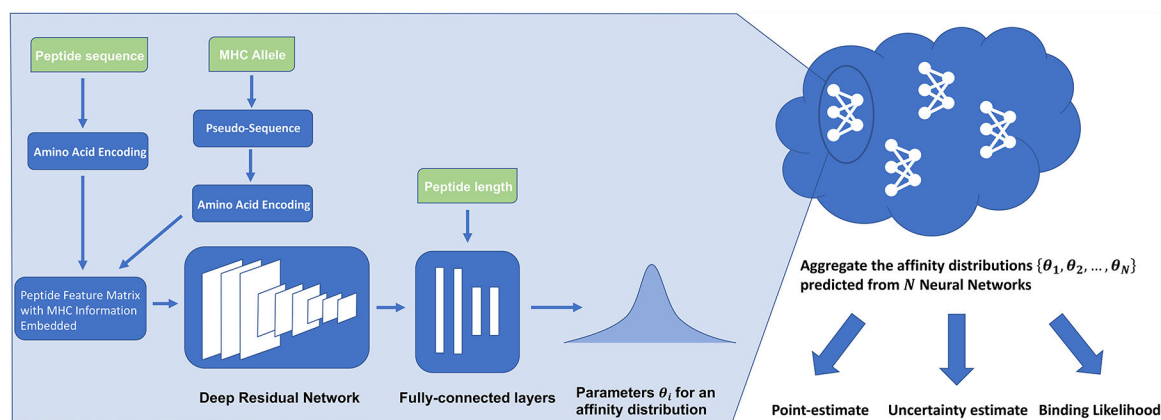


Figure 1.

PUFFIN is an ensemble of neural networks. MHC alleles are represented by a pseudo-sequence of annotated amino acids in contact with the peptide. Each amino acid in the peptide sequence and the MHC pseudo-sequence is encoded with both a one-hot amino acid sequence encoding and the row of BLOSUM50 that corresponds to the amino acid, referred to as “Amino Acid Encoding” in the diagram. The amino acid encoding of the peptide and MHC are combined to form a feature matrix. Given the encoded feature matrix and the peptide length as input, each network in the ensemble outputs the parameters of a peptide-MHC affinity distribution (Beta distribution for class I MHC and Normal distribution for class II MHC). The ensemble-wide average of the affinity mean is a point estimate of affinity. The epistemic uncertainty is the ensemble-wide variance of the affinity mean and the aleatoric uncertainty is the ensemble-wide mean of the affinity variance. For a given affinity threshold T , the binding likelihood of a peptide-MHC pair is defined as the probability that the affinity exceeds T under a Beta (for class I MHC) or Normal (for class II MHC) distribution with parameters averaged across the ensemble.

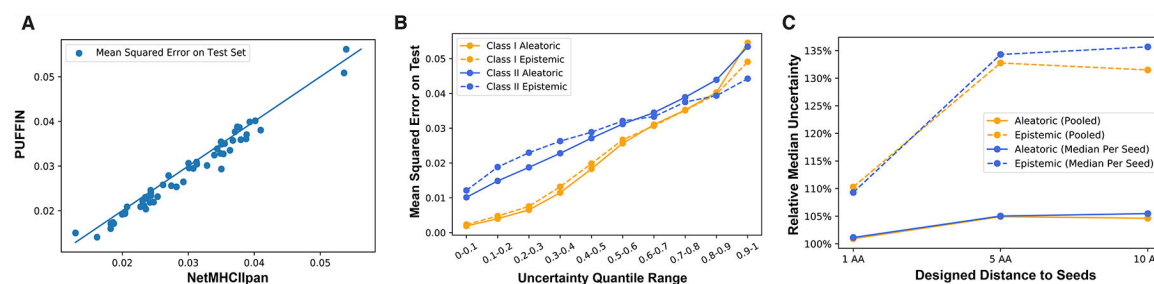


Figure 2.

PUFFIN provides state-of-the-art affinity prediction as well as reliable uncertainty metrics for each prediction. (A) The mean squared error of PUFFIN (y-axis) and NetMHCIIpan (x-axis) over 55 class II MHC alleles. (B) The x-axis shows the 10-quantiles of PUFFIN's aleatoric (solid line) and epistemic (dash line) uncertainty estimates for class I (yellow) and class II (blue) MHC. The y-axis shows the mean squared error on the held-out examples with an uncertainty estimate in the corresponding quantile. (C) The x-axis denotes three sets of derived sequences with different designed distance to the training examples. The y-axis denotes the median epistemic (dash line) and aleatoric (solid line) uncertainty of the corresponding set of derived sequences normalized against the median uncertainty of the seed. Yellow lines represent the results from jointly considering derived sequences mutated from all seeds. Blue lines represent the results from only including the median uncertainty of derived sequences mutated from the same seed in the calculation.

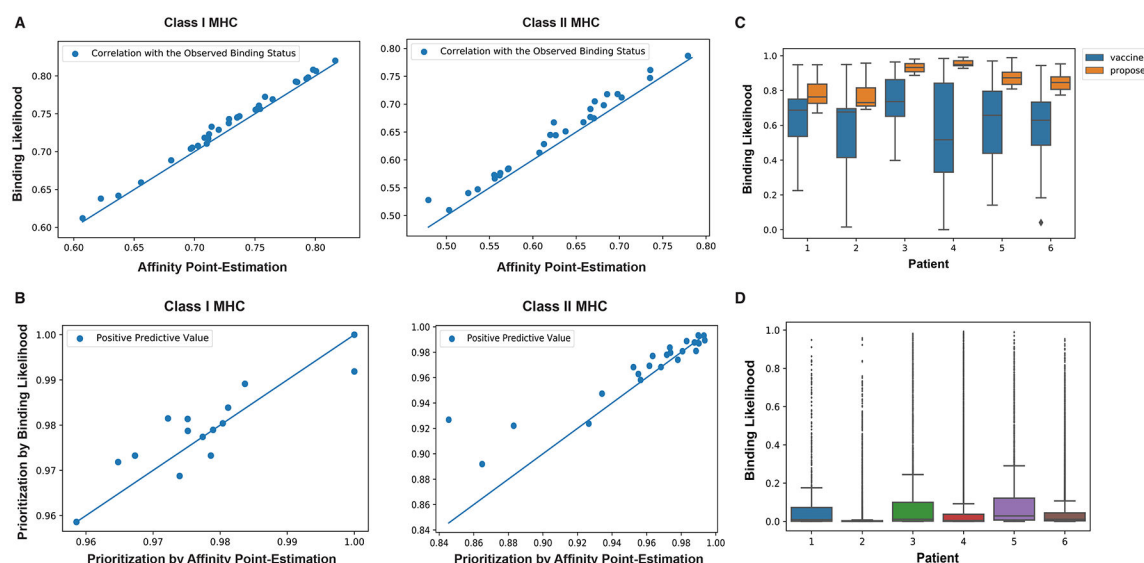


Figure 3.

Uncertainty estimation improves the design of high-affinity peptides. (A) The Point-Biserial correlation with the observed binding status binarized with respect to an affinity cutoff of 500 nM for both PUFFIN's binding likelihood (y-axis) and affinity point estimate (x-axis). Each point denotes an allele of class I (left) or class II (right) MHC. The diagonal line denotes cases where the performance is the same for both metrics. (B) The positive predictive value for the peptides with high binding likelihoods (>95%) (y-axis) and for the peptides with the highest predicted affinity (x-axis). Each point denotes an allele of class I (left) or class II (right) MHC. The diagonal line denotes cases where PPV is the same for both approaches. (C) Binding likelihoods of peptides in published neoantigen vaccines (blue) and an alternative set of mutation-spanning peptides prioritized by binding likelihood (orange). Boxplots show median (the center line in the box), 25th and 75th percentiles (the boundaries of the box), 1.5 interquartile range (the ends of the whiskers), and outliers (points outside of the whiskers). Same applies to (D). (D) Binding likelihoods of all mutation-spanning peptides for each patient.

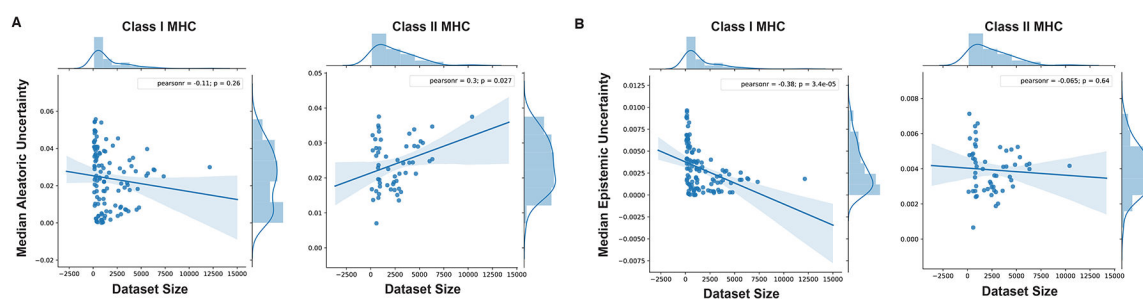


Figure 4.

IEDB power by allele and MHC class. (A) The correlation between dataset size and the median aleatoric uncertainty for all class I (left) and class II (right) MHC alleles available in IEDB. (B) The correlation between dataset size and the median epistemic uncertainty for all class I (left) and class II (right) MHC alleles available in IEDB.

Table 1.

Comparison of performance in predicting class II MHC-peptide binding affinity. For auROC, F1 Score and Point-Biserial Correlation, positive samples are defined as ones with an affinity stronger than 500 nM. PUFFIN-mean denotes using PUFFIN's "mean affinity" as prediction. PUFFIN-BL denotes using PUFFIN's "binding likelihood" as prediction. NetMHCIIpan + PUFFIN-mean denotes using the average of the affinity predictions from NetMHCIIpan and PUFFIN-mean. For PUFFIN-BL, the prediction is no longer affinity and thus F1 score (which defines positive predictions using an affinity cutoff), mean squared error and R^2 are not meaningful metrics. The best performing method for each metric is highlighted in bold.

	auROC	F1 Score	Mean Squared Error	R^2	Spearman Correlation	Point-Biserial Correlation
NetMHCIIpan	0.8727	0.7415	0.03056	0.5529	0.7367	0.6301
PUFFIN-mean	0.8774	0.7504	0.02956	0.5679	0.7450	0.6381
PUFFIN-BL	0.8795	N/A	N/A	N/A	0.7424	0.6561
NetMHCIIpan + PUFFIN-mean	0.8808	0.7516	0.02904	0.5751	0.7516	0.6427

Table 2.

Comparison of performance in predicting class I MHC-peptide binding affinity. For auROC, F1 Score and Point-Biserial Correlation, positive samples are defined as ones with an affinity stronger than 500 nM. PUFFIN-mean denotes using PUFFIN's "mean affinity" as prediction. PUFFIN-BL denotes using PUFFIN's "binding likelihood" as prediction. For PUFFIN-BL, the prediction is no longer affinity and thus F1 score, which defines positive predictions using an affinity cutoff, is not a meaningful metric. The performance for MHCnuggets, NetMHCpan, and MHCflurry in this table was reported in Bhattacharya et al., in which the Point-Biserial correlations were not used as a metric. The best performing method for each metric is highlighted in bold.

Model	auROC	Kendall's tau	F1 Score	Point-Biserial Correlation
MHCnuggets	0.931	0.589	0.810	-
NetMHCpan	0.933	0.584	0.803	-
MHCflurry	0.933	0.587	0.785	-
PUFFIN-mean	0.935	0.599	0.802	0.756
PUFFIN-BL	0.936	0.573	N/A	0.767

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Bacterial and Virus Strains		
Biological Samples		
Chemicals, Peptides, and Recombinant Proteins		
Critical Commercial Assays		
Deposited Data		
Class I MHC binding affinity data in IEDB	Nielsen et al., 2016	http://www.cbs.dtu.dk/services/NetMHCpan-3.0/
Class II MHC binding affinity data in IEDB	Jensen et al., 2018	http://www.cbs.dtu.dk/services/NetMHCIIpan-3.2/
Curated class I MHC benchmark dataset	Bhattacharya et al., 2017; Mendeley Data	https://www.biorxiv.org/content/10.1101/154757v2 ; https://doi.org/10.17632/jwhmrdx268.1
Experimental Models: Cell Lines		
Experimental Models: Organisms/Strains		
Oligonucleotides		
Recombinant DNA		

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
PUFFIN	This paper	http://github.com/gifford-lab/PUFFIN
NNAlign2.0	Nielsen et al., 2017	http://www.cbs.dtu.dk/services/NNAlign-2.0/
Pytorch	Facebook	https://pytorch.org/
Other		
BLOSUM50 matrix	National Center for Biotechnology Information (NCBI)	https://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/data/BLOSUM50