

Constrained Submodular Optimization for Vaccine Design

Zheng Dai, David K. Gifford

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
zhengdai@mit.edu, gifford@mit.edu

Abstract

Advances in machine learning have enabled the prediction of immune system responses to prophylactic and therapeutic vaccines. However, the engineering task of designing vaccines remains a challenge. In particular, the genetic variability of the human immune system makes it difficult to design peptide vaccines that provide widespread immunity in vaccinated populations. We introduce a framework for evaluating and designing peptide vaccines that uses probabilistic machine learning models, and demonstrate its ability to produce designs for a SARS-CoV-2 vaccine that outperform previous designs. We provide a theoretical analysis of the approximability, scalability, and complexity of our framework.

Introduction

Peptide vaccines that expand and activate T cells have emerged as a promising prophylactic and therapeutic approach for addressing health related challenges including infectious diseases and cancer (Malonis, Lai, and Vergnolle 2019). In contrast to more conventional live-attenuated vaccines that are based on entire organisms, or subunit vaccines that are based on entire protein subunits, peptide vaccines are based on a small set of protein fragments (peptides) that are sufficient to induce a T cell immune response, enabling the elicitation of far more targeted responses that avoid allergic and reactogenic responses (Li et al. 2014).

The design of a peptide vaccine consists of selecting immunogenic protein fragments, usually referred to as epitopes (Li et al. 2014), that when included in a vaccine expand epitope specific T cells. Advances in machine learning have enabled our ability to predict which peptides will be presented by major histocompatibility complex (MHC) molecules for surveillance by the adaptive immune system (Ching et al. 2018; Reynisson et al. 2020), which can be used to identify which epitopes will be displayed (Sohail et al. 2021).

The epitopes displayed by an individual depend upon the specific alleles of their MHC genes, and thus the peptides displayed by the immune system can vary greatly from individual to individual (Zaitouna, Kaur, and Raghavan 2020). Therefore, the engineering task of finding a set of peptides that is predicted to be displayed by a large portion of the pop-

ulation remains challenging despite progress on the peptide-MHC display task.

In this work we introduce a framework for evaluating and designing peptide vaccines that uses probabilistic interpretations of machine learning models, and demonstrate its ability to produce designs for the SARS-CoV-2 vaccine design task that outperform previous designs. We complement this with a theoretical analysis of the approximability, scalability, and complexity of our framework, which may be of independent interest.

Our Contribution

To improve the effectiveness of a vaccine it is important to introduce redundancies into its design so the failure of a single displayed peptide to elicit an immune response does not become a single point of failure (Liu et al. 2022). Vaccines designed with an n -times coverage objective aim to obtain at least n immunogenic peptide “hits” in each person. Having more than one “hit” provides redundancy to expand multiple T cell clonotypes in an individual to fight disease, protects against peptide sequence drift resulting from pathogen or tumor mutations, protects against the loss of an MHC gene, and accounts for the variability of peptide immunogenicity between individuals.

We show that optimizing the population coverage for strict n -times coverage guarantees cannot be tractably approximated to any constant factor assuming the intractability of the GAP-SMALL-SET EXPANSION problem (Raghavendra and Steurer 2010), a result which may be of independent interest. We therefore propose a *diminishing returns* framework that uses a soft redundancy guarantee as its objective. The resulting objective is both submodular and monotonic, and can therefore be approximated via a greedy approach which we call *Optivax-P*. We supplement the theoretical improvement with an empirical comparison of vaccines designed using our approach and previous designs. Our proposed framework also contributes the following desirable properties: it makes explicit the utility of having redundancy, does not discount the benefits of being covered without redundancy, and is able to reason with uncertainty. We demonstrate how uncertainty values for epitope identification can be derived by calibrating state-of-the-art peptide-MHC display predictors.

While redundancies in a design are important, it is also

important that they be *dissimilar redundancies*, since reasons for failure may be shared between similar peptides. This additional constraint that selected peptides be dissimilar is problematic as it allows the problem formulation to encode NP-hard graph problems that in general cannot be approximated to any constant factor. However, by parameterizing on the structure of the constraints, we can derive lower bounds for the performance of the greedy approach which show that the greedy approach can still provide approximation guarantees under certain assumptions. These bounds may also be of independent interest.

Related Work

The use of computational methods to aid vaccine design has taken on an increasingly important role in the vaccine design process over the past two decades (Moise et al. 2015). Much of the advancement stems from improvements in the epitope identification task, which has seen impressive improvements with advances in data collection strategies and machine learning (Ching et al. 2018; Reynisson et al. 2020). While good epitope prediction tools are essential to vaccine design, the focus of this work is on the downstream task of calibrating the predictions and selecting defined epitopes for vaccine inclusion.

Earlier works on vaccine design are reviewed in Oyarzun and Kobe (2015), and employ discrete optimization techniques such as integer linear programming and genetic algorithms to optimize population coverage. However, they do not anticipate or solve the problem of coverage with dissimilar redundancies (Liu et al. 2022), which we do in this work. Furthermore, they do not consider the epistemic uncertainty associated with epitope predictions, which we do.

Our work is closely related to the work in (Liu et al. 2022), where the use of an objective that accounts for dissimilar redundancies is proposed. However, approximating their proposed objective to any constant factor appears to be an intractable problem, while our objective permits constant factor approximations in polynomial time. Our framework also allows for reasoning about *redundancies* with uncertainty, which theirs does not.

Presentation

In the remaining four sections of the paper we present the optimization problem that we wish to solve (A Diminishing Returns Objective Enables Theoretical Performance Guarantees), provide an algorithm for solving the problem and analyze its runtime and approximation guarantees (Methods), apply our framework to the SARS-CoV-2 vaccine design problem (Results), and conclude with a discussion (Discussion). Theorems are presented where appropriate throughout. Proofs, including intuitive descriptions, are relegated to Appendix A for improved flow (see Data Availability regarding appendices).

A Diminishing Returns Objective Enables Theoretical Performance Guarantees

Our goal in this section is to formalize the vaccine design problem as an optimization problem. We first show theoret-

ical barriers to obtaining performance guarantees for previous formalizations, and then introduce the *diminishing returns* framework which addresses this.

Peptide vaccines are designed by considering the peptide sequence(s) of a target of interest, for example the proteome of a virus, and selecting a small set of peptides within the target sequences to include in the vaccine. Vaccine peptides are selected such that they elicit an immune response in a large portion of a susceptible population that we wish to vaccinate. This is done by selecting vaccine peptides that are displayed on the cell surface by MHC proteins. The resulting peptide-MHC complexes activate the cellular immune system. The challenge of selecting a set of peptides arises from the polymorphism present in MHCs within a population. Different MHC alleles have different peptide binding properties, so the peptides must be carefully chosen in order to elicit widespread immune responses from a given population.

Preliminaries

Let $\mathbb{R}^{\geq 0}$ denote non-negative real numbers. Let S be some finite set of elements. Let $F : 2^S \rightarrow \mathbb{R}^{\geq 0}$. We say F is submodular if $F(S_1 \cup \{e\}) - F(S_1) \geq F(S_2 \cup \{e\}) - F(S_2)$ whenever $S_1 \subseteq S_2$ and $e \in S$, and we say that F is monotonically increasing if $S_1 \subseteq S_2 \implies F(S_1) \leq F(S_2)$ for all $S_1, S_2 \subseteq S$.

Suppose $G = (V, E)$ is a graph. For simplicity, we will at times use G_V to denote its vertex set V and G_E to denote its edge set E . The k th power of G , denoted G^k , is defined as the graph (G_V^k, G_E^k) , where $G_V^k = G_V$ and G_E^k contains all pairs of vertices between which there exists a path of length less than or equal to k in G .

We will use $\mathbb{1}_X$ to denote an indicator that evaluates to 1 if X is true and 0 otherwise for any proposition X .

Optimizing Population Coverage with Redundancies Is Computationally Difficult

It is important for a vaccine to cause the display of multiple epitopes in individuals to provide redundancy in the activation of T cell clonotypes, to expand multiple T cell clonotypes in an individual to fight disease, to protect against peptide sequence drift as a consequence of pathogen or tumor mutations, to protect against the loss of an MHC gene, and to account for the variability of peptide immunogenicity between individuals (Liu et al. 2022). In Liu et al. (2022), the authors showed that previous vaccine designs fail to cover significant portions of the population when coverage criteria include these redundancies. To address this, they introduce the n -times coverage framework, which involves solving the max n -times coverage problem. The problem is defined as follows:

Definition 1. *Given a ground set, a set of weights over the ground set, a collection of multisets whose elements are from the ground set, and some cardinality constraint k , find a collection of k multisets such that the aggregate weights of the elements in the ground set that are covered at least n times is maximized.*

The sum of the weights of the elements that are covered at least n times is then called the n -times coverage. For vaccine design, the ground set corresponds to MHC genotypes, the weights correspond to the percentage of the population with the genotypes, and each multiset corresponds to a peptide, which covers certain genotypes a variable number of times. Solving this problem with cardinality constraint k then gives a vaccine design consisting of k peptides, with the objective that a large portion of the population display at least n peptides (i.e. have at least n peptide-MHC hits).

While this is a natural extension of earlier vaccine design paradigms that do not account for redundancies, it is a computationally difficult problem. The authors have shown in their work that this is an NP-hard optimization problem, and so they propose heuristic approaches. However their proposed approaches have no performance guarantees. Here, we show that this problem is related to GAP-SMALL-SET EXPANSION, which suggests that finding any constant factor approximation cannot be achieved in polynomial time.

Theorem 1. *For any $\epsilon > 0$, if there exists a polynomial time algorithm that can achieve an approximation factor of ϵ to max n -times coverage, then there exists a polynomial time algorithm that can decide GAP-SMALL-SET EXPANSION(η) for some $\eta \in (0, 0.5)$.*

There is currently no known polynomial time algorithm for GAP-SMALL-SET EXPANSION(η) for any $\eta \in (0, 0.5)$. The *Small Set Expansion Hypothesis* conjectures that GAP-SMALL-SET EXPANSION(η) is NP-hard for any $\eta \in (0, 0.5)$, and is currently an open problem related to the Unique Games Conjecture (Raghavendra and Steurer 2010).

A Diminishing Returns Framework for Vaccine Design Provides a Submodular Optimization Objective

A key reason underlying the complexity of the max n -times coverage problem is that the utility of a peptide may be hidden until we are close to reaching n -times coverage. This makes it difficult select peptides optimally before its utility becomes apparent. To address this, we propose a diminishing returns framework, where peptides will improve the objective at any coverage level. Intuitively, this provides a “gradient” along which an optimization procedure can climb.

Formally, let $U : \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}^{\geq 0}$ be some non-negative monotonically increasing concave function such that $U(0) = 0$. Let \mathcal{M} denote the set of MHC genotypes observed in the population. Let $w : \mathcal{M} \rightarrow \mathbb{R}^{\geq 0}$ be a weight function that gives the frequency of each genotype in the population. Let \mathcal{P} denote the set of candidate peptides from which a subset is selected for the vaccine design. If $p \in \mathcal{P}$ and $m \in \mathcal{M}$, let $display(p, m)$ denote the predicate of whether p is displayed in an individual with genotype m . To model uncertainty, we let $display(p, m)$ vary over the sample space of some probability space, and we assume that the subset of the probability space where $display(p, m)$ evaluates to true is always measurable. The objective, parametrized by U , can then be written as follows:

$$\mathcal{F}_U(S) = \sum_{m \in \mathcal{M}} w(m) \mathbb{E}[U(\sum_{p \in S} \mathbb{1}_{display(p, m)})] \quad (1)$$

Where $S \subseteq \mathcal{P}$ is the set of peptides selected for vaccine inclusion. This objective is a monotonically increasing submodular function.

Theorem 2. *For any $U : \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}^{\geq 0}$ that is monotonically increasing and concave, \mathcal{F}_U is a monotonically increasing submodular function.*

As a consequence, we can attain a $(1 - e^{-1})$ -factor approximation using the greedy approach if no additional constraints are given aside from the cardinality of the peptide set (Nemhauser, Wolsey, and Fisher 1978). Beyond submodularity, this objective contributes the following desirable properties: first, it accounts for the fact that having peptide-MHC hits is useful, even if the redundancy does not obtain a given threshold. While having high redundancy is better than having low redundancy, having low redundancy is better than not displaying any peptides. Second, the utility we expect from attaining a given number of peptide-MHC hits is made explicit through U . Third, it allows reasoning with uncertainty by allowing $display(p, m)$ to be an uncertain event. Many prediction models output a soft classification instead of a hard one, which we can calibrate to attach uncertainties to the classifications.

For an arbitrary distribution over the set of indicator variables $\mathbb{1}_{display(p, m)}$ we may need to approximate the expectation in \mathcal{F}_U via sampling. However, we can compute the objective \mathcal{F}_U exactly and efficiently if we suppose that for a given MHC genotype m , the set of indicator variables $\mathbb{1}_{display(p, m)}$ are independent. This is almost certainly false given a sufficiently large pool of peptide sequences, since we should be able to significantly improve the performance of a predictor by training it on a sufficiently large number of peptide sequences. However, we can weaken this assumption to k -wise independence if we only consider vaccine designs that include at most k peptides. For values of k that are reasonable in the context of designing peptide vaccines, this assumption is more reasonable than the full independence assumption.

Under the independence assumption we can calculate the objective by computing the distribution of the sum via iterated convolutions of Bernoulli distributions, and then taking the expectation using the distribution (see Appendix C for additional details). This runs in time $\mathcal{O}(|\mathcal{M}||S|^2)$, where $|\mathcal{M}|$ is the number of genotypes and $|S|$ is the number of peptides in the vaccine design.

Peptide Selections Need to Be Constrained to Avoid Unreasonable Designs

We impose two types of constraints on the set of peptides selected for vaccine inclusion: a cardinality constraint, and a set of pairwise constraints.

The cardinality constraint is necessary since our objective function is monotonically increasing. Therefore, the full set of candidate peptides \mathcal{P} will maximize it. This is undesirable, since peptide vaccines need to be compact to permit

effective delivery and to induce effective intolerance in the context of limited immune system capacity. Therefore, we will impose a cardinality constraint on the set of selected peptides such that it cannot exceed a given size k .

The pairwise constraints are required to avoid very similar peptides from being included. Peptide candidates for vaccine inclusion are generated by sliding windows of various sizes across the protein sequence we wish to target. This produces peptides that are highly similar in sequence, such as nested sequences, and including highly related sequences does not truly improve the effectiveness of the vaccine. Furthermore, the assumption that the variables indicating peptide-MHC interactions are independent likely does not hold when peptides are very similar, since it is possible that the predictor makes use of similar features, which result in systematic errors. Therefore, we introduce a set of pairwise constraints \mathcal{G} as a graph where the vertex set $\mathcal{G}_V = \mathcal{P}$, and where edges exist between peptides that are deemed redundant. We then require that the peptides in the vaccine design form an independent set within \mathcal{G} .

Methods

A Greedy Approach Provides Performance Guarantees under the Diminishing Returns Framework

Our goal is the following: given a peptide set \mathcal{P} , a set of MHC genotypes \mathcal{M} , binding credences between all peptides and MHCs, a monotonically increasing concave utility function $U : \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}^{\geq 0}$ with $U(0) = 0$, a cardinality constraint k , and pairwise constraints \mathcal{G} , find a set $S \subseteq \mathcal{P}$ that satisfies all the constraints and maximizes the objective function $\mathcal{F}_U(S)$. We define the binding credence between a peptide p and an MHC m as the measure of the subset of the probability space where $\text{display}(p, m)$ evaluates to true. Practically, these values behave like probabilities. We use the term credence to emphasize the epistemic nature of the uncertainty.

We present Optimax-P, a greedy approach outlined in Algorithm 1, to produce a solution to this problem. The procedure is straightforward: at each iteration we add the peptide that maximally improves the solution to the solution set, then eliminate that peptide and all similar peptides from consideration for all future steps.

Runtime Analysis of Optimax-P The naive runtime is $\mathcal{O}(k^3|\mathcal{P}||\mathcal{M}|)$: the objective function is evaluated $|\mathcal{P}|$ times to compute the $\arg \max$, the $\arg \max$ is computed at most k times, and each evaluation of the objective function takes time $\mathcal{O}(k^2|\mathcal{M}|)$ since the designs will never contain more than k elements.

We can improve the runtime by evaluating the marginal improvement rather than the full objective, bringing the overall runtime down to $\mathcal{O}(k^2|\mathcal{P}||\mathcal{M}|)$ (see Appendix C). We can further vectorize the computation to evaluate the $\arg \max$ in $\mathcal{O}(1)$ vector operations, reducing the runtime to $\mathcal{O}(k)$ vector operations and $\mathcal{O}(|\mathcal{P}|k)$ operations for constraint handling (see Appendix C). However, vector operations require batching if $|\mathcal{M}|$ and $|\mathcal{P}|$ are large, so those parameters still play a significant role in the runtime.

Algorithm 1: Optimax-P

Input: A ground set of candidate peptides \mathcal{P} , a cardinality constraint k , a similarity graph (\mathcal{P}, E) , and a monotone submodular function $F : 2^{\mathcal{P}} \rightarrow \mathbb{R}^{\geq 0}$ where $F(\emptyset) = 0$
Output: A set $S \subseteq \mathcal{P}$ such that $|S| \leq k$

```

 $S \leftarrow \emptyset$ 
 $Q \leftarrow \mathcal{P}$ 
while  $(Q \neq \emptyset) \wedge (|S| < k)$  do
   $x \leftarrow \arg \max_{x \in Q} F(S \cup \{x\})$ 
   $S \leftarrow S \cup x$ 
   $N \leftarrow \{y | \{x, y\} \in E\} \cup \{x\}$ 
   $Q \leftarrow Q \setminus N$ 
end while
return  $S$ 

```

Our implementation can generate designs of size $k \approx 10^2$ over a peptide set of size $|\mathcal{P}| \approx 10^3$ with $|\mathcal{M}| \approx 10^6$ genotypes in approximately 5 minutes when parallelized over 8 Titan RTX GPUs. See Appendix C for additional details.

Approximation Ratio of Optimax-P Let S^* denote the true optimum of the optimization problem. If we let each peptide and MHC genotype interact with probability 1 and let k be sufficiently large, then the desired optimization is equivalent to finding the maximum independent set within \mathcal{G} . Finding any constant factor approximation to max-clique is NP-hard (Zuckerman 2006), which then immediately implies that S^* cannot be approximated to any constant factor. Therefore, the quality of the solution produced by Optimax-P cannot be unconditionally bounded by a constant factor with respect to S^* , since Optimax-P runs in polynomial time.

However, we can bound the solution by looking at the graph structure of \mathcal{G} . Since we are considering cases where similarity relations are mostly generated from sliding windows over linear sequences, we might expect the resulting graph to be of low degree. Let $\Delta(\mathcal{G})$ denote the degree of \mathcal{G} . Note that in the special case where $\Delta(\mathcal{G}) = 0$, there are no pairwise constraints, so the problem reduces to the optimization of a monotonic submodular function under a cardinality constraint. It is well established that the greedy approach attains an approximation ratio of $(1 - e^{-1})$ in this case (Nemhauser, Wolsey, and Fisher 1978), and that attaining an approximation ratio of $(1 - e^{-1} + \epsilon)$ for any $\epsilon > 0$ is NP-hard (Feige 1998).

Another property we can look at is the graph power of \mathcal{G} . We may expect that in the case where a graph looks like a path, taking the graph power would not add too many extra constraints, in which case replacing \mathcal{G} with its graph power \mathcal{G}^p would not yield a optimum that is too different. Let S_p^* denote the solutions to the more constrained optimizations:

$$S_p^* = \arg \max_{\substack{S \subseteq \mathcal{P}: |S| \leq k \\ v_1, v_2 \in S \implies \{v_1, v_2\} \notin \mathcal{G}_E^p}} \mathcal{F}_U(S) \quad (2)$$

We can then bound the output of Optimax-P by incorporating these extra graph parameters:

Theorem 3. *Let \hat{S} be the output of Optimax-P. Then:*

1. If $\Delta(\mathcal{G}) = 0$, then $\mathcal{F}_U(\hat{S}) \geq \mathcal{F}_U(S^*)(1 - e^{-1})$
2. If $\Delta(\mathcal{G}) > 0$, then $\mathcal{F}_U(\hat{S}) \geq \max(\frac{\mathcal{F}_U(S_2^*)}{2}, \frac{\mathcal{F}_U(S^*)}{1+\Delta(\mathcal{G})})$

We can upper bound the best possible performance of polynomial time algorithms.

Theorem 4. *Unless $P = NP$, there exists no polynomial time algorithm that can output an approximation \hat{S} that guarantees either of the following on all inputs for any $\epsilon > 0$:*

1. $\mathcal{F}_U(\hat{S}) \geq \mathcal{F}_U(S_2^*)(1 - e^{-1} + \epsilon)$
2. $\mathcal{F}_U(\hat{S}) \geq \mathcal{F}_U(S^*)(\frac{1}{1+\Delta(\mathcal{G})})^{(1-\epsilon)}$

Using Machine Learning Based Models to Attain Binding Credences

We define $display(p, m)$ as a random event in a probability space of beliefs that peptide p is presented by MHC genotype m . We use well established state-of-the-art neural network based models (NetMHCpan4.1 and NetMHCpanII4.0 (Reynisson et al. 2020)) to generate predictions that we use to derive $\Pr(display(p, m))$. We will assume that the derived beliefs are independent (although it is sufficient to assume k -wise independence, and only between events that share a genotype - see prior section titled “A Diminishing Returns Framework for Vaccine Design Provides a Submodular Optimization Objective”). While this may not necessarily be true for closely related sequences, we circumvent this by constraining our designs so that they do not contain closely related sequences (see prior section titled “Peptide Selections Need to Be Constrained to Avoid Unreasonable Designs”).

There are two classes of MHC molecules that we need to design our vaccine to bind to. However, due to molecular differences between the two molecules the peptides they bind are largely disjoint. We can therefore simplify the problem to that of producing two separate designs, one for Class I MHCs and one for Class II MHCs. NetMHCpan4.1 provides predictions between peptides and Class I MHCs while NetMHCIIpan4.0 provides predictions between peptides and Class II MHCs. For simplicity, we will refer to these predictors together as NetMHCpan.

Having fixed our class of interest, we then define an MHC genotype to be a set of 3-6 distinct alleles. NetMHCpan provides binding predictions between peptides and individual alleles rather than genotypes, so to attain credences for whether a peptide is displayed by a genotype we assume independence between our beliefs that the peptides are displayed by individual alleles and compute the credence as the following:

$$\Pr(display(p, m)) = 1 - \prod_{x \in m} (1 - c_{p,x}) \quad (3)$$

Where $c_{p,x}$ is the credence for binding between a peptide p and a specific allele x .

NetMHCpan outputs likelihoods for binding between any given peptide p and any given MHC molecule x . However, it is important to ensure that this likelihood is well calibrated. By default, NetMHCpan calibrates itself by comparing the scores it outputs against a repertoire of naturally occurring

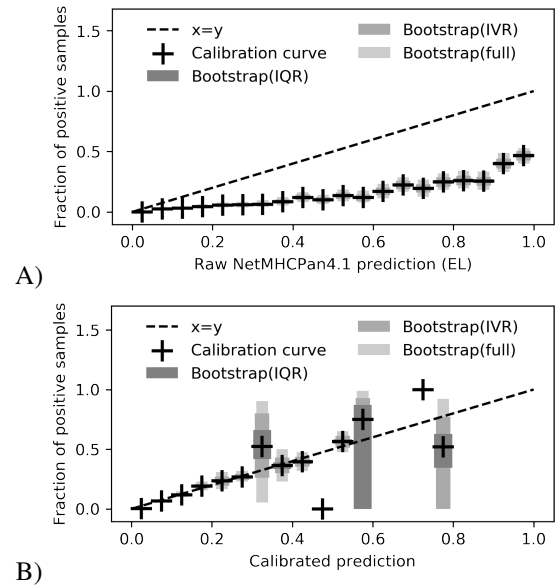


Figure 1: Calibration curves for Class I MHC with uncalibrated (A) and calibrated (B) predictions. Matching plots for Class II MHC can be found in Appendix B. The curve is made of 20 equally spaced bins between 0 and 1. Populated bins and the fraction of positive samples they contain are indicated by a “+”, with the surrounding column indicating the interquartile (IQR), interventile (IVR) and full range of a set of 1000 bootstrapped values.

peptides and classifying a sequence as being displayed if it scores higher than 99.5% of those peptides when characterizing binding to Class I MHC and 98% of those peptides when classifying binding to Class II MHC.

To attain well calibrated credences, we make use of publicly available datasets that were used to validate NetMHCpan, which contain no overlap with the dataset used to train NetMHCpan (Reynisson et al. 2020). The dataset consists of a set of epitopes, the MHC molecule they bind to, and the natural context they occur in. The epitopes were used as positive samples while the context sequences were used to generate negative samples. The samples were weighted such that the ratio of the weight of all positive samples to that of all negative samples is 1 : 199 for Class I MHC and 1 : 49 for Class II MHC to match the fraction of natural peptides that NetMHCpan implicitly assumes to be binders. The samples were then all fed into NetMHCpan to produce predictions. See Appendix B for additional details. The samples were then binned into 20 equally sized bins and the weighted fraction of positive samples within each bin was calculated to produce a calibration curve, which we present in Figure 1A. We then generate a calibration function \mathcal{H} by minimizing the following objective:

$$\sum_{i=1}^{n-n'} \left(\frac{\sum_{j=1}^{n'} w_{i+j} (\mathcal{H}(x_{i+j}) - y_{i+j})}{\sum_{j=1}^{n'} w_{i+j}} \right)^2 \quad (4)$$

Where all samples are indexed by an integer between 1

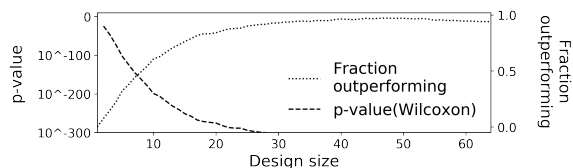


Figure 2: For each design size between 1 and 64 inclusive, we compute the fraction of the 2000 settings under which Optimav-P outperforms both baselines, and test the hypothesis of whether Optimav-P outperforms the best score obtained by the baselines via a one-sided Wilcoxon signed-rank test. The fraction and p-values (corrected for multiple hypothesis by a factor of 64) are then plotted as a function of the number of peptides selected.

and n , x_i denotes the NetMHCpan predicted value of the sample, y_i is 1 if the sample is positive and 0 if the sample is negative, and w_i is the weight assigned to the sample. The samples are ordered such that $x_i \leq x_{i+1}$ for all $0 \leq i < n$, so n' can be interpreted as the window size. We constrain \mathcal{H} to be non-decreasing and non-negative, so this is a form of isotonic regression. Additional details can be found in Appendix B.

If NetMHCpan outputs y on an input peptide p and MHC m , we set our credence that p binds to m as $\mathcal{H}(y)$. A calibration curve of the calibrated predictions of the validation dataset is shown in Figure 1B.

Results

Greedy Selection Outperforms Baseline Methods

Optimav-P has good worst case theoretical guarantees which are even optimal in the absence of pairwise constraints unless $P=NP$. Here we empirically check its performance against two baseline approaches in random settings. The first baseline is the *random approach*, where peptides are chosen randomly. The second baseline is via *linear approximation*, where the concave function U in Equation 1 is removed to produce a surrogate objective. This surrogate is linear, so it can be trivially optimized to produce a solution.

We compare the algorithms by generating peptide vaccine designs of 64 different sizes in 2000 randomly generated settings, where each setting randomizes the number of genotypes and peptides, the display credences, and the concave function U in the objective \mathcal{F}_U (Equation 1). Details can be found in Appendix D. We find that Optimav-P outperforms the baselines in most settings (Figure 2).

Designing a SARS-CoV2 Vaccine with Optimav-P

We apply Optimav-P in a practical setting by producing vaccine designs for SARS-CoV2. To design our vaccine, we use a set of candidate peptides sourced from Liu et al. (2022) which includes peptides from SARS-CoV2 that have been filtered for undesirable properties like high mutation rates, cleavage, and glycosylation. Peptides that were present in the human proteome were also removed, since they may trigger adverse autoimmune responses. Liu et al. (2022)

have also published a set of genotypes and their frequencies which we use. These frequencies are derived from diverse populations and have been selected to be representative of the global population. We calculated binding credences for each peptide-genotype pair as described in the section titled “Using Machine Learning Based Models to Attain Binding Credences”. Additionally, if a peptide was not present in representative genomes of the Omicron BA.1 and Omicron BA.2 variants of SARS-CoV2 (accession number OM873778 and OW123901 respectively, both retrieved from the The COVID-19 Data Portal (Harrison et al. 2021)) then we set the credence of that peptide being displayed on any MHC molecule to 0.

We applied Optimav-P to design vaccines with peptide sets of size 1 and 150 inclusive. Designs were constrained such that no pair of peptides can be within 3 edits (insertions, deletions, or substitutions) of each other for the MHC Class I design, and 5 edits for the MHC Class II design. Designs were optimized for the objective \mathcal{F}_{U_T} for T between 1 and 20 inclusive, where U_T is defined as:

$$U_T(x) = \min(x, T) \quad (5)$$

T can be viewed as a threshold parameter. This corresponds to a model where each peptide-MHC hit provides incremental protection until a person attains T hits, at which point they are fully protected and stop seeing additional benefits. A comparison between Optimav-P designs and previous designs is shown in Figure 3, where we can see all Optimav-P designs score substantially higher on \mathcal{F}_{U_5} . Evaluations on \mathcal{F}_{U_T} for other T show similar trends and are presented in Appendix D.

While this demonstrates that our optimization procedure works and that there are deficiencies in previous designs that are addressed through our designs if our utility model is reflective of reality, it may not be entirely surprising that vaccine designs optimized for \mathcal{F}_{U_T} score well on \mathcal{F}_{U_T} , even for mismatching T . To control for this, we also evaluate designs produced by Optimav-P on the n -times coverage objective proposed by Liu et al. (2022) and described in the section titled “Optimizing Population Coverage with Redundancies Is Computationally Difficult”. We modify our credences such that they are only 0 or 1, and such that they closely match the values used by Liu et al. (2022) (i.e. $\Pr(\text{display}(p, m)) = 1$ if and only if the genotype m is present within the peptide p , where peptides are viewed as multisets of genotypes in the max n -times coverage framework).

We then produced designs of size 19 using U_T for all T between 1 and 10 inclusive. These designs were then assigned a score equal to their n -times coverage for n between 1 and 20 inclusive. In Figure 4 we compare these scores to the scores attained by designs generated by Liu et al. (2022), which also each contain 19 peptides. We see that Optimav-P is highly competitive even when evaluated on an objective separate from the one it was optimized for.

Discussion

We have introduced the diminishing returns framework and Optimav-P for designing epitope based peptide vaccines. The

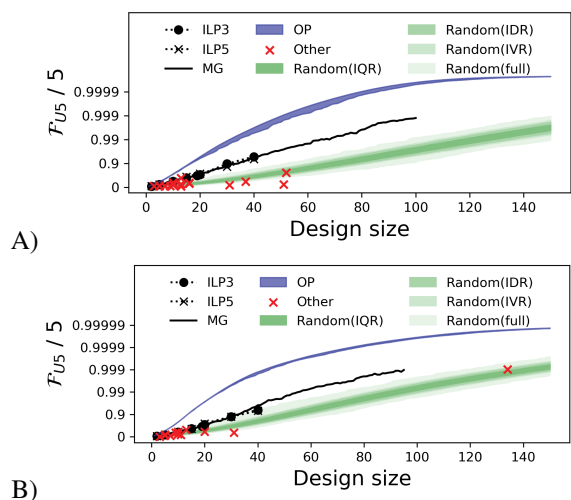


Figure 3: For each vaccine design S we compute $\mathcal{F}_{U_5}(S)$, divide it by 5, and plot it as a function of $|S|$. Designs for both MHC Class I (A) and MHC Class II (B) are given. For each design size we plot the entire range of designs of that size generated using Optivax-P optimized for \mathcal{F}_{U_T} with $1 \leq T \leq 20$ (OP), and the interquartile (IQR), interdecile (IDR), interventile (IVR), and full range of 1000 designs of that size sampled uniformly at random. We also plot the locations of the ILP $n=3$ (ILP3), ILP $n=5$ (ILP5), and MarginalGreedy (MG) designs from Liu et al. (2022) as well as 29 other designs (Other) (Abdelmageed et al. 2020; Ahmed, Quadeer, and McKay 2020; Nazneen Akhand et al. 2020; Alam et al. 2020; Banerjee, Santra, and Maiti 2020; Baruah and Bose 2020; Bhattacharya et al. 2020; Fast, Altman, and Chen 2020; Gupta, Mishra, and Niraj 2020; Herst et al. 2020; Lee and Koohy 2020; Mitra et al. 2020; Poran et al. 2020; Ramaiah and Arumugaswami 2021; Saha and Prasad 2020; Singh et al. 2020; Srivastava et al. 2020; Tahir ul Qamar et al. 2020; Vashi, Jagrit, and Kumar 2020).

framework is based on constrained submodular optimization, which permits Optivax-P to provide performance guarantees despite the NP-hardness of the problem, unlike previous approaches. We also show how we can probabilistically interpret the outputs of machine learning models to allow reasoning with uncertainty within the framework. Finally, we demonstrated that Optivax-P achieves superior performance against past vaccine designs on the SARS-CoV-2 vaccine design task, and achieves comparable performance even when evaluated against previous objectives.

Optivax-P is highly scalable, allowing us to optimize over potentially millions of candidate peptides for a single vaccine. The ability to reason with uncertainty also gives us a much richer language for expressing the properties of a peptide: for instance, instead of filtering out peptides prone to mutation, we can consider the probability of sequence drift. These factors allow us to consider a far wider range of peptide vaccine design tasks, which we plan to use in the future.

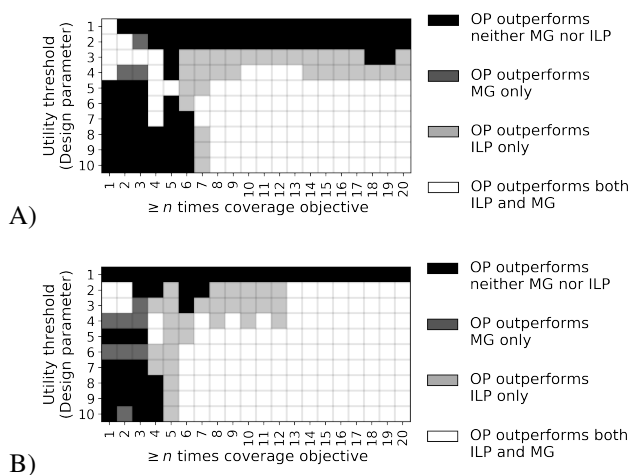


Figure 4: We compare designs generated by Optivax-P (OP) to designs generated by the ILP and MarginalGreedy (MG) approaches from Liu et al. (2022) using the n -times coverage objective. A white square at position (x, y) indicates that the design proposed by Optivax-P optimized against \mathcal{F}_{U_y} outperforms both the ILP and MarginalGreedy design when evaluated with x -times coverage. A black square indicates that the Optivax-P design was outperformed by the ILP or MG design, and a gray square indicates that the Optivax-P design was outperformed by either the ILP or MG design. Comparisons for both MHC Class I (A) and MHC Class II (B) are given.

Data Availability

Code, data, appendices, and demo are accessible from <https://gifford-lab.github.io/DiminishingReturns>.

Acknowledgements

This work was supported in part by Schmidt Futures and a C3.ai grant. The authors would also like to thank Ge Liu, Alexander Dimitrakakis, Brandon Carter, and members of the Gifford lab for useful discussions and feedback.

References

- Abdelmageed, M. I.; Abdelmoneim, A. H.; Mustafa, M. I.; Elfadol, N. M.; Murshed, N. S.; Shantier, S. W.; and Makhawi, A. M. 2020. Design of multi epitope-based peptide vaccine against E protein of human COVID-19: An immunoinformatics approach. *bioRxiv*, (2020.02.04.934232).
- Ahmed, S. F.; Quadeer, A. A.; and McKay, M. R. 2020. Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses*, 12(3): 254.
- Alam, A.; Khan, A.; Imam, N.; Siddiqui, M. F.; Waseem, M.; Malik, M. Z.; and Ishrat, R. 2020. Design of an Epitope-Based Peptide Vaccine against the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2): A Vaccine-informatics Approach. *bioRxiv*, (2020.05.03.074930).

- Banerjee, A.; Santra, D.; and Maiti, S. 2020. Energetics based epitope screening in SARS CoV-2 (COVID 19) spike glycoprotein by Immuno-informatic analysis aiming to a suitable vaccine development. *bioRxiv*, (2020.04.02.021725).
- Baruah, V.; and Bose, S. 2020. Immunoinformatics-aided identification of T cell and B cell epitopes in the surface glycoprotein of 2019-nCoV. *Journal of medical virology*, 92(5): 495–500.
- Bhattacharya, M.; Sharma, A. R.; Patra, P.; Ghosh, P.; Sharma, G.; Patra, B. C.; Lee, S.-S.; and Chakraborty, C. 2020. Development of epitope-based peptide vaccine against novel coronavirus 2019 (SARS-COV-2): Immunoinformatics approach. *Journal of medical virology*, 92(6): 618–631.
- Ching, T.; Himmelstein, D. S.; Beaulieu-Jones, B. K.; Kalinin, A. A.; Do, B. T.; Way, G. P.; Ferrero, E.; Agapow, P.-M.; Zietz, M.; Hoffman, M. M.; et al. 2018. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141): 20170387.
- Fast, E.; Altman, R. B.; and Chen, B. 2020. Potential T-cell and B-cell Epitopes of 2019-nCoV. *bioRxiv*, (2020.02.19.955484).
- Feige, U. 1998. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM (JACM)*, 45(4): 634–652.
- Gupta, E.; Mishra, R. K.; and Niraj, R. R. K. 2020. Identification of potential vaccine candidates against SARS-CoV-2, A step forward to fight COVID-19: A Reverse Vaccinology Approach. *bioRxiv*, (2020.04.13.039198).
- Harrison, P. W.; Lopez, R.; Rahman, N.; Allen, S. G.; Aslam, R.; Buso, N.; Cummins, C.; Fathy, Y.; Felix, E.; Glont, M.; et al. 2021. The COVID-19 Data Portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. *Nucleic acids research*, 49(W1): W619–W623.
- Herst, C. V.; Burkholz, S.; Sidney, J.; Sette, A.; Harris, P. E.; Massey, S.; Brasel, T.; Cunha-Neto, E.; Rosa, D. S.; Chao, W. C. H.; et al. 2020. An effective CTL peptide vaccine for Ebola Zaire Based on Survivors' CD8+ targeting of a particular nucleocapsid protein epitope with potential implications for COVID-19 vaccine design. *Vaccine*, 38(28): 4464–4475.
- Lee, C. H.; and Koohy, H. 2020. In silico identification of vaccine targets for 2019-nCoV. *F1000Research*, 9: 145.
- Li, W.; Joshi, M. D.; Singhanian, S.; Ramsey, K. H.; and Murthy, A. K. 2014. Peptide vaccine: progress and challenges. *Vaccines*, 2(3): 515–536.
- Liu, G.; Dimitrakakis, A.; Carter, B.; and Gifford, D. K. 2022. Maximum n-times Coverage for Vaccine Design. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 4153. OpenReview.net.
- Malonis, R. J.; Lai, J. R.; and Vergnolle, O. 2019. Peptide-based vaccines: current progress and future challenges. *Chemical reviews*, 120(6): 3210–3229.
- Mitra, D.; Pandey, J.; Jain, A.; and Swaroop, S. 2020. In silico design of Multi-epitope-based peptide vaccine against SARS-CoV-2 using its spike protein. *bioRxiv*, (2020.04.23.055467).
- Moise, L.; Gutierrez, A.; Kibria, F.; Martin, R.; Tassone, R.; Liu, R.; Terry, F.; Martin, B.; and De Groot, A. S. 2015. iVAX: An integrated toolkit for the selection and optimization of antigens and the design of epitope-driven vaccines. *Human vaccines & immunotherapeutics*, 11(9): 2312–2321.
- Nazneen Akhand, M. R.; Azim, K. F.; Hoque, S. F.; Moli, M. A.; Joy, B. D.; Akter, H.; Afif, I. K.; Ahmed, N.; and Hasan, M. 2020. Genome based Evolutionary study of SARS-CoV-2 towards the Prediction of Epitope Based Chimeric Vaccine. *bioRxiv*, (2020.04.15.036285).
- Nemhauser, G. L.; Wolsey, L. A.; and Fisher, M. L. 1978. An analysis of approximations for maximizing submodular set functions—I. *Mathematical programming*, 14(1): 265–294.
- Oyarzun, P.; and Kobe, B. 2015. Computer-aided design of T-cell epitope-based vaccines: addressing population coverage. *International journal of immunogenetics*, 42(5): 313–321.
- Poran, A.; Harjanto, D.; Malloy, M.; Rooney, M. S.; Srinivasan, L.; and Gaynor, R. B. 2020. Sequence-based prediction of vaccine targets for inducing T cell responses to SARS-CoV-2 utilizing the bioinformatics predictor RECON. *bioRxiv*, (2020.04.06.027805).
- Raghavendra, P.; and Steurer, D. 2010. Graph expansion and the unique games conjecture. In *Proceedings of the forty-second ACM symposium on Theory of computing*, 755–764.
- Ramaiah, A.; and Arumugaswami, V. 2021. Insights into Cross-species Evolution of Novel Human Coronavirus SARS-CoV-2 and Defining Immune Determinants for Vaccine Development. *bioRxiv*, (2020.01.29.925867).
- Reynisson, B.; Alvarez, B.; Paul, S.; Peters, B.; and Nielsen, M. 2020. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic acids research*, 48(W1): W449–W454.
- Saha, R.; and Prasad, B. V. L. S. 2020. In silico approach for designing of a multi-epitope based vaccine against novel Coronavirus (SARS-COV-2). *bioRxiv*, (2020.03.31.017459).
- Singh, A.; Thakur, M.; Sharma, L. K.; and Chandra, K. 2020. Designing a multi-epitope peptide based vaccine against SARS-CoV-2. *Scientific reports*, 10(1): 1–12.
- Sohail, M. S.; Ahmed, S. F.; Quadeer, A. A.; and McKay, M. R. 2021. In silico T cell epitope identification for SARS-CoV-2: Progress and perspectives. *Advanced drug delivery reviews*, 171: 29–47.
- Srivastava, S.; Verma, S.; Kamthania, M.; Kaur, R.; Badyal, R. K.; Saxena, A. K.; Shin, H.-J.; Kolbe, M.; and Pandey, K. C. 2020. Structural basis to design multi-epitope vaccines against Novel Coronavirus 19 (COVID19) infection, the on-going pandemic emergency: an in silico approach. *bioRxiv*, (2020.04.01.019299).
- Tahir ul Qamar, M.; Rehman, A.; Tusleem, K.; Ashfaq, U. A.; Qasim, M.; Zhu, X.; Fatima, I.; Shahid, F.; and Chen, L.-L. 2020. Designing of a next generation multi-epitope based vaccine (MEV) against SARS-COV-2: Immunoinformatics and in silico approaches. *PloS one*, 15(12): e0244176.

Vashi, Y.; Jagrit, V.; and Kumar, S. 2020. Understanding the B and T cell epitopes of spike protein of severe acute respiratory syndrome coronavirus-2: A computational way to predict the immunogens. *Infection, Genetics and Evolution*, 84: 104382.

Zaitouna, A. J.; Kaur, A.; and Raghavan, M. 2020. Variations in MHC class I antigen presentation and immunopeptidome selection pathways. *F1000Research*, 9: 1177.

Zuckerman, D. 2006. Linear degree extractors and the inapproximability of max clique and chromatic number. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, 681–690.