Bioinformatics, 36(26), 2020, i643–i650
doi: 10.1093/bioinformatics/btaa790
ECCB2020

Genomes

Joint epitope selection and spacer design for string-of-beads vaccines

Emilio Dorigatti[1,2,*] and Benjamin Schubert[2,3]

[1]Faculty of Mathematics Informatics and Statistics, Ludwig Maximilian Universität, München 80333, Germany, [2]Institute of Computational Biology, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg 85764, Germany and [3]Department of Mathematics, Technical University of Munich, Garching bei München 85748, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: Conceptually, epitope-based vaccine design poses two distinct problems: (i) selecting the best epitopes to elicit the strongest possible immune response and (ii) arranging and linking them through short spacer sequences to string-of-beads vaccines, so that their recovery likelihood during antigen processing is maximized. Current state-of-the-art approaches solve this design problem sequentially. Consequently, such approaches are unable to capture the inter-dependencies between the two design steps, usually emphasizing theoretical immunogenicity over correct vaccine processing, thus resulting in vaccines with less effective immunogenicity *in vivo*.

Results: In this work, we present a computational approach based on linear programming, called JessEV, that solves both design steps simultaneously, allowing to weigh the selection of a set of epitopes that have great immunogenic potential against their assembly into a string-of-beads construct that provides a high chance of recovery. We conducted Monte Carlo cleavage simulations to show that a fixed set of epitopes often cannot be assembled adequately, whereas selecting epitopes to accommodate proper cleavage requirements substantially improves their recovery probability and thus the effective immunogenicity, pathogen and population coverage of the resulting vaccines by at least 2-fold.

Availability and implementation: The software and the data analyzed are available at https://github.com/SchubertLab/JessEV.

Contact: edo@stat.uni-muenchen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

One of the most prominent approaches to rational vaccine design against cancer (Hu *et al.*, 2018; Ott *et al.*, 2017; Sahin and Türeci, 2018) and infectious diseases (Audran *et al.*, 2005; Barouch *et al.*, 2018) are so-called epitope-based vaccines (EVs). EVs consist of short immunogenic peptides, called epitopes, that are presented on human leukocyte antigen (HLA) molecules and elicit a T-cell response. Such vaccines can be produced quickly and cheaply with proven technologies and easily preserved. They also eliminate the risk of reversion to virulence present in regular attenuated vaccines, and can be engineered to reduce potential toxicity and inflammatory reactions (Liu, 2019).

The design process of EVs is composed of three stages: discovery of potential epitopes, selection of a subset to be included in the vaccine and their arrangement into a vaccine. As it has become clear that delivery of mixtures of separate epitopes is not effective in inducing a strong immune response, delivery strategies have been developed that assemble the selected epitopes into concatenated polypeptide vaccines, so-called string-of-beads vaccines, thereby considerably increasing their immunogenicity (Yang *et al.*, 1996). In a string-of-beads construct, the epitopes are linked by short sequences of few amino acids, called spacers, designed to elicit correct proteasomal cleavage at the N- and C-termini of the epitopes. This increases the recovery likelihood of the epitopes and the effective immunogenicity of the vaccine as a whole.

Current state-of-the-art methods for string-of-beads design approach epitope selection and vaccine assembly independently. First, epitopes are selected to maximize the theoretical immunogenicity of the vaccine subject to additional design constraints (Lundegaard *et al.*, 2010; Toussaint *et al.*, 2008) but completely disregard the subsequent assembly and processing of the vaccine. Only in a second step, the selected epitopes are assembled into a string-of-beads vaccine optimizing their recovery likelihood either using predetermined, hand-designed spacer sequences (Velders *et al.*, 2001) or spacers specifically optimized for each epitope pair (Schubert and Kohlbacher, 2016).

Although the immunogenicity of the selected epitopes and the cleavage likelihood of the assembled vaccine strongly influence each other (Cornet *et al.*, 2006), existing approaches cannot adequately

capture and exploit this trade-off due to their sequential nature. This often leads to string-of-beads vaccines with theoretically high immunogenicity but undesirable cleavage patterns that prevent many epitopes from being recovered during vaccine processing fundamentally reducing the vaccine's effective immunogenicity.

Our main contribution is, therefore, an approach that considers these two steps together using mixed-integer linear programming (MILP). Our mathematical framework is able to select and assemble a subset of maximally immunogenic epitopes that conform to prespecified design constraints regarding their conservation, coverage of pathogens and HLA alleles, as well as cleavage probabilities of their interior, and N- and C-termini. Through extensive Monte Carlo cleavage simulations, we show that the resulting vaccines provide much greater epitope recovery rates compared to vaccines designed with a sequential approach. As a consequence, the effective immunogenicity, pathogen and population coverage are significantly increased, demonstrating the necessity of modeling both design steps simultaneously.

2 Materials and methods

2.1 A unifying framework for epitope selection and assembly of string-of-bead vaccines

An epitope is effective only if it is recovered from the vaccine polypeptide (i.e. cleavage occurs at its termini and not in its interior). Hence, epitopes should not only be selected based on their theoretical immunogenicity but also based on their proteasomal cleavage likelihood and therefore their recovery probability. By controlling the cleavage likelihood through optimized arrangement of epitopes and specifically designing spacer sequences, we can increase the probability that the epitopes of a vaccine are recovered correctly. Other quantities, besides cleavage and immunogenicity, such as coverage of HLA and pathogenic variability, as well as epitope conservation might be of interest as well to make the vaccine robust and broadly applicable.

Therefore, the design problem can be described as finding the optimal set of epitopes $E \subset \mathcal{E}$ of given size k that maximizes the immunogenicity $I(E)$ while conforming with other pre-specified design criteria, and simultaneously assembling the epitopes into a string-of-beads vaccine with spacer sequences S_{ij} for each pair of connected epitopes $(e_i, e_j) \in E \times E$ that maximizes their cleavage likelihood.

We formulate this optimization problem as an MILP, which guarantees a globally optimal string-of-beads vaccine. The MILP is conceptually divided into two blocks. The base linear program (Table 1) contains the basic constraints needed to encode the vaccine design problem, ensuring consistency of the resulting solution, reconstructing the amino acid sequence of the selected epitopes and spacers and computing cleavage scores for each position. The cleavage score is proportional to the cleavage probability in that specific position, and is computed as a sum of offset-dependent scores of the surrounding amino acids. As we allow spacers of variable length, it is not possible to directly calculate the offsets needed to query these cleavage contributions. Instead, we reformulate the cleavage calculation by linearizing a bivariate function $\mathbb{Z} \times \mathbb{N} \to \mathbb{R}$ mapping offset and amino acid to their individual score contribution.

The second building block contains optional constraints related to the selection of epitopes in the vaccine and to bound cleavage scores in certain locations of the string-of-beads construct. The epitope selection constraints force the vaccine to cover a given minimum amount of pathogens and/or HLA alleles. Furthermore, they can restrict the epitopes selected to have a certain minimum average conservation. As most epitopes have extremely low conservation, we found it preferable to focus on the average, rather than the minimum. The cleavage constraints are applied to certain critical locations: the N- and C-termini of the epitopes, their interior and the interior of the spacers. We will later give suggestions of broadly applicable values for the cleavage site thresholds. A full description of the MILP can be found in Supplementary Table S1.

2.2 Immunogenicity model

As in Toussaint *et al.* (2008), we define the overall contribution of an epitope e to the vaccine immunogenicity as the weighted average of the log-transformed HLA binding strengths ι_{ea} over a specified set of HLA alleles $\mathcal{A}$:

$$I(e) := \sum_{a \in \mathcal{A}} p_a \cdot (1 - \log_{50k} \iota_{ea}) \tag{1}$$

where p_a is the probability of the allele a occurring in an individual of the target population. The set $\mathcal{A}$ of alleles has to be chosen carefully beforehand to target a specific population, or to match the patient's genotype in case of personalized therapy. We chose HLA binding affinity as proxy of immunogenicity as it has been shown to be strongly correlated with epitope immunogenicity (Paul *et al.*, 2013; Sette *et al.*, 1994). A variety of models that predict HLA binding affinity with high accuracy have been developed (Peters *et al.*, 2020). We chose NetMHCpan (Jurtz *et al.*, 2017) as it is considered state-of-the-art and most widely used; however, the framework is agnostic to the specific immunogenicity predictor, and the results presented here are expected to hold for any HLA binding affinity model.

The overall immunogenicity of the vaccine is then the sum of the individual immunogenicities of the chosen epitopes weighted by the HLA frequencies of a target population, and constitutes the objective of the MILP (OBJ in Table 1).

Table 1. The base linear program that selects epitopes and spacers (consistency constraints), reconstructs the amino acid sequence (not shown), and computes the cleavage score for each position of the sequence (cleavage computation constraints and PSSM access constraints)

Maximize		
(OBJ)	$\sum_{e \in \mathcal{E}} \sum_{p \in \mathcal{P}} x_{ep} \cdot I(e)$	
Subject to: (consistency constraints)		
(C1)	$\sum_{e \in \mathcal{E}} x_{ep} = 1$	$\forall p \in \mathcal{P}$
(C2)	$\sum_{p \in \mathcal{P}} x_{ep} \leq 1$	$\forall e \in \mathcal{E}$
(C3)	$\sum_{a \in \mathcal{A}} y_{ats} \leq 1$	$\forall s \in \mathcal{S}, t \in \mathcal{T}$
(C4)	$x_{ep}, y_{aqs} \in \{0,1\}$	$\forall a, e, p, q, s$
Subject to: (cleavage computation constraints)		
(C5)	$\sum_{a \in \mathcal{A}} s_{pa} = f_p$	$\forall p \in \mathcal{Q}$
(C6)	$\sum_{r=p+1}^{q} f_r = o_{pq}$	$\forall p, q \in \mathcal{Q} : p \leq q$
(C7)	$-\sum_{r=q}^{p-1} f_r = o_{pq}$	$\forall p, q \in \mathcal{Q} : p > q$
(C8)	$f_p \cdot \sum_{q \in \mathcal{P}} p_{pq} = c_p$	$\forall p \in \mathcal{Q}$
(C9)	$f_p \in \{0,1\}$	$\forall p \in \mathcal{Q}$
Subject to: (PSSM access constraints)		
(C10)	$\sum_{i=1}^{m} \sum_{j \in \mathcal{A}} \phi_{ij} \lambda_{pqi} s_{qj} = p_{pq}$	$\forall p, q \in \mathcal{Q}$
(C11)	$\lambda_{pq0} o_{pq} + \sum_{i=1}^{6} \mathrm{O}_i \lambda_{pqi} = o_{pq}$	$\forall p, q \in \mathcal{Q}$
(C12)	$\sum_{i=1}^{n} \lambda_{pqi} = \alpha_{pq} \beta_{pq}$	$\forall p, q \in \mathcal{Q}$
(C13)	$\lambda_{pq0} = 1 - \alpha_{pq} \beta_{pq}$	$\forall p, q \in \mathcal{Q}$
(C14)	$o_{pq} - (L+5) \cdot \alpha_{pq} \leq -4.5$	$\forall p, q \in \mathcal{Q}$
(C15)	$o_{pq} + (L+2) \cdot \beta_{pq} \geq 1.5$	$\forall p, q \in \mathcal{Q}$
(C16)	$\alpha_{pq}, \beta_{pq}, \lambda_{pqi} \in \{0,1\}$	$\forall p, q \in \mathcal{Q}$ $0 \leq i \leq 6$ $1 \leq j \leq 20$

where $\mathcal{A}, \mathcal{E}$ and $\mathcal{P}$, indices of amino acids, epitopes and epitope positions; $\mathcal{Q}, \mathcal{S}$ and $\mathcal{T}$, indices for sequence positions, spacers and positions inside spacers; $I(e)$, the immunogenicity of epitope e; x_{ep}, equals one if epitope e is in position p of the vaccine; y_{ats}, equals one if amino acid a is in position t of spacer s; s_{pa}, equals one if amino acid a is in position s of the whole sequence (computation not shown in this table); ϕ_{ij}, content of the PSSM for amino acid A_j at offset O_i. Zero if i is out of bounds; L, maximum length of the vaccine sequence.

2.3 Cleavage site model

Through data-driven models, we can assign a proteasomal cleavage probability to each position within a peptide sequence. State-of-the-art cleavage site prediction models take the amino acids of neighboring positions into account and assume their influence to be independent (Dönnes and Kohlbacher, 2005; Kuttler *et al.*, 2000; Li *et al.*, 2012; Tenzer *et al.*, 2005). Using observational data, these models estimate $p(A_o = a_{k+o}|C_k = 1)$, the probability that an amino acid o positions away from the cleavage site k is a_{k+o}, and $p(A_o = a_{k+o})$, the probability of amino acid a_{k+o} occurring in a protein. Assuming independence, the cleavage probability is then expressed as:

$$\frac{p(C_k = 1|a_{k-N_c}, \ldots, a_{k+N_t})}{p(C_k = 1)} = \exp\Big(\sum_{o=-N_c}^{N_t} \phi(o, a_{k+o})\Big) \tag{2}$$

where $\phi(o, a_{k+o})$ is the content of a position-specific scoring matrix (PSSM) for amino acid a_{k+o} at offset o from the cleavage point k, and represents the log-ratio of the probabilities which are multiplied in Equation (2). These models implicitly assume that $C_{k-N_c} = \cdots = C_{k+N_t} = 0$. Therefore, we complement Equation (2) with the following additional condition:

$$p(C_k = 1|C_{k+o} = 1) = 0 \quad \forall o \in \{-N_c, \ldots, N_t\}. \tag{3}$$

Note that the resulting score is still relative to the prior probability $p(C_k = 1) = p_c$ of cleavage, which may vary according to the host organism. Given that the average length of the peptides cleaved by the proteasome is between seven and nine amino acids (Nussbaum *et al.*, 1998), a reasonable value for this prior probability could be between 0.15 and 0.20, but as it is not clear how to set this parameter, we will investigate in detail its influence on the results. Here, we used the Proteasomal Cleavage Matrix (PCM), a position-specific scoring matrix (PSSM) proposed by Dönnes and Kohlbacher (2005) that uses four C-terminal amino acids ($N_c = 4$) and two N-terminal amino acids ($N_t = 1$) to predict a cleavage site. It has been shown to give robust and generalizable predictions.

Given this model, the recovery event of an epitope e can then be computed as:

$$R_e = C_{\mathrm{Nt}(e)} \cdot C_{\mathrm{Ct}(e)} \cdot \prod_{p \in \mathrm{In}(e)} (1 - C_p) \tag{4}$$

where $\mathrm{Nt}(e)$ and $\mathrm{Ct}(e)$ are the positions of e's termini and $\mathrm{In}(e)$ are the residues inside e.

2.4 Linearizing PSSM indexing

The difficulty in querying a PSSM within the specified MILP arises from the necessity of dynamically calculating the indexing position due to the variable length of each spacer sequence. We solve this issue by bounding the spacer length from above and below, which gives us a fixed reference frame in which we can specify the amino acid sequence of the spacer while allowing some position to be empty.

Formally, the cleavage score c_p at position p can be computed as follows:

$$c_p = f_p \cdot \sum_{q=1}^{L} (f_q \cdot \phi(o_{pq}, a_q)) \tag{5}$$

where $f_k = 1$ if there is an amino acid in position k, L is the maximum length of the vaccine sequence, $\phi(o, a)$ is the entry of the PSSM or zero if out of bounds, a_q is the amino acid in position q and o_{pq} is the number of amino acids between positions p and q including the one at q. Equation (5) corresponds to constraint C8 in Table 1.

The position-specific indicators f can be computed easily from the position-amino acid indicators (Table 1 C5) and the offset o_{pq} can be computed as the sum of the indicators for the positions between p (not included) and q (included). As o_{pq} is used to index the PSSM, we define o_{pq} to be negative when $q < p$ and use two constraints for the positive and negative cases (Table 1 C6 and C7, respectively).

Indexing into the PSSM can then be expressed as:

$$\phi(o_{pq}, a_q) = \begin{cases} \phi_{ij} & \text{if } o_{pq} = O_i \wedge a_q = A_j \\ 0 & \text{if } L \leq o < -4 \\ 0 & \text{if } 1 < o \leq U \end{cases} \tag{6}$$

with the pivots $O_i \in \{-4, \ldots, 1\}$ and $A_j \in \{1, \ldots, 20\}$ indicating the offset and the amino acid, and ϕ_{ij} the entry of the PSSM. We also require lower and upper bounds L and U for o; the maximum length of the sequence suffices. All the quantities involved in Equation (6) are integers, except for ϕ_{ij} which is a real number.

Equation (6) can be linearized similarly to how piece-wise linear functions are (Vielma *et al.*, 2010), with a few adaptations for our specific case. In particular, we associate an indicator variable to each pivot, $s_{qj} = 1[a_q = A_j]$ and $\lambda_{pqi} = 1[o_{pq} = O_i]$, and retrieve the cleavage score from the PSSM as a linear combination of these indicators with the respective pivot in constraint C10 (Table 1).

The indicators s_{qj} can be computed easily from x and y (Supplementary Table S1). The appropriate λ can be computed by comparing every pivot O_i to the actual offset o_{pq} (Table 1 C11), but require a default value of zero if o_{pq} is out of the bounds of the PSSM (i.e. -4 and 1). To this end, we introduce a new indicator λ_{pq0} that is not used to compute cleavage. Constraint C11 (Table 1) can always be satisfied by choosing $\lambda_{pq0} = 1$. Therefore, further constraints were added to force this to happen only if the offset is actually out of the PSSM bounds. Consequently, we introduce two additional indicator variables $\alpha_{pq} = 1[o_{pq} > 1]$ (Table 1 C14) and $\beta_{pq} = 1[o_{pq} < -4]$ (Table 1 C15), and set $\lambda_{pq0} = 1 - \alpha_{pq}\beta_{pq}$ (Table 1 C13).

2.5 Monte Carlo simulations

Given the cleavage scores, we are interested in estimating the probability that an epitope e of the string-of-beads vaccine is recovered, which happens when $C_{\mathrm{Nt}(e)} = C_{\mathrm{Ct}(e)} = 1$ and $C_p = 0$ for $p \in \mathrm{In}(e)$ (i.e. cleavage happens at e's termini and not inside it). Equation (2) defines the probability of cleavage at a certain position conditioned on the surrounding amino acids. To calculate C_k for each position within the string-of-beads vaccine, we assume that the proteasome cleaves the vaccine from N- to C-terminus and define $p(C_k = 1|a_{k-4}, \ldots, a_{k+1}) = 0$ if any of $C_{k-4}, \ldots, C_{k-1}$ have been cleaved before (i.e. $\exists C_{k-i} = 1, i \in [1, 4]$). In this case, we assume that $C_k = 0$ instead. Based on this, we can define the cleavage event as a stochastic process indexed by the position k in the sequence:

$$C_k = \begin{cases} 0 & \text{if } k < 5 \vee \sum_{i=-4}^{-1} C_{k+i} > 0 \\ 1 & \text{with probability } p(C_k = 1|a_{k-4}, \ldots, a_{k-1}) \end{cases} \tag{7}$$

We can estimate the recovery probability of each epitope $p(R_e = 1)$ by sampling from this stochastic process through Monte Carlo simulations and computing the ratio of successful recoveries, as defined in Equation (4), over the number of simulations performed.

Using the epitope recovery probability, we can then estimate quantities of interest such as the effective immunogenicity and effective coverage of the vaccine as the expectation of the respective metric under the recovery probabilities. For example, the effective immunogenicity is computed as:

$$\mathbb{E}[I(E)] = \sum_{e \in E} \mathbb{E}[I(e)] = \sum_{e \in E} I(e) \cdot p(R_e = 1) \tag{8}$$

We use these simulations to evaluate the vaccines after they have been designed by solving the linear program.

2.6 Dataset

Nine-mer epitopes were extracted from 275 randomly selected sequences of the Nef gene of HIV-1 subtypes B and C, downloaded from the HIV Sequence Database (Foley *et al.*, 2018; Los Alamos National Laboratory, 2019), for a total of 13 668 epitopes, whose

binding affinity was then predicted by NetMHCpan (Jurtz *et al.*, 2017). We considered the same 27 HLA alleles and their frequencies as Toussaint *et al.* (2011), that together provide a theoretical coverage of 91.3% of the world population.

2.7 Implementation

The software was implemented in Python (van Rossum, 2001), using Pyomo (Hart *et al.*, 2011, 2017) to formulate the linear program and Gurobi (Gurobi Optimization, 2020) to solve it. The implementation of OptiTope (Toussaint *et al.*, 2008) for epitope selection and OptiVac (Schubert and Kohlbacher, 2016) for spacer design of the sequential approach was provided by FRED2 (Schubert *et al.*, 2016). NumPy (van der Walt *et al.*, 2011), Scipy (Virtanen *et al.*, 2020), Pandas (McKinney, 2010), statsmodels (Seabold and Perktold, 2010), Matplotlib (Hunter, 2007) and Seaborn (Waskom *et al.*, 2017) were used to analyze and visualize the results in the IPython environment (Pérez and Granger, 2007).

3 Results

Our vaccines were compared against vaccines designed by first selecting the optimal set of epitopes using OptiTope (Toussaint *et al.*, 2008), then finding the optimal ordering and spacer sequences using OptiVac (Schubert and Kohlbacher, 2016). We refer to this procedure as the 'sequentia' approach/design, and the method we proposed is referred to as 'JessEV'. Due to the large number of experiments performed, we limited the length of all vaccines to five epitopes and at most four-amino acids spacers.

3.1 Smaller cleavage likelihood inside epitopes and larger cleavage likelihood at their termini is possible

We created 30 sets of 5000 epitopes extracted without replacement from the complete set of 13 668 epitopes, and designed a vaccine for each set with the sequential approach and with JessEV.

Both approaches could reach similar cleavage likelihoods at the epitope junction sites (Fig. 2), yet the sequentially designed vaccines often exhibited less favorable cleavage patterns within epitopes (Fig. 2a). As the sequential epitope selection method cannot consider vaccine processing during epitope selection, the subsequent epitope assembly model has limited opportunity to generate a favorable cleavage pattern. Even though only 4% of epitope residues had larger cleavage than a terminal residue, 44% of them had a score larger than zero, i.e. cleavage was more likely than the prior in those locations. This led to frequent cleavage within epitopes, which would nullify their therapeutic effected *in vivo*.

With our framework, we could take this into account by enforcing negative cleavage score inside the epitopes, scores larger than two at the termini, and smaller than negative two inside the spacers, corresponding to 7.4 times more/less likely than the prior cleavage likelihood, respectively. This caused the average score of residues inside the epitopes to decrease significantly (unpaired *t*-test, $t = -9.40$, $P - \text{value} = 9 \times 10^{-21}$) from −0.40 for the sequential design to −1.01 for JessEV, for an effect size of −0.50. Similarly, the average scores at the termini significantly increased ($t = 16.36$, $P - \text{value} = 1 \times 10^{-47}$) from 1.70 to 2.16, for an effect size of 1.17. The largest difference was inside the spacers, where the average score decreased from −0.25 to −5.22 ($t = -47.74$, $P - \text{value} = 9 \times 10^{-220}$, effect size of −3.62).

We then used NetChop Cterm (Nielsen *et al.*, 2005) to obtain an independent prediction of the cleavage sites for every bootstrap, and counted how many cleavage events happened in the termini, epitopes and spacers (Fig. 2b). On average, there were, respectively, 1.73, 9.40 and 6.07 cleavage events for sequentially designed vaccines, and 4.37, 6.81 and 6.44 for JessEV's designs. The effect sizes were 1.90, −1.99 and 0.37. We then fitted a Poisson regression model considering the choice of algorithmic design as the independent variable, and found that the difference between the number of cleavage events inside the spacers was not significant (0.06 ± 1.06, $z = 0.57$ and $P - \text{value} = 0.57$), but the difference for termini and epitopes were (0.92 ± 0.17, $z = 5.56$, $P - \text{value} = 3 \times 10^{-8}$ and -0.32 ± 0.09, $z = -3.39$, $P - \text{value} = 7 \times 10^{-4}$).

Next, we used 1000 Monte Carlo cleavage simulations with a prior probability of 0.15 to estimate the effective immunogenicity of the vaccines. With this prior probability, the 275 source sequences were cleaved in fragments of average length 9.08 amino acids, almost identical to the length of the typical HLA class I-bound

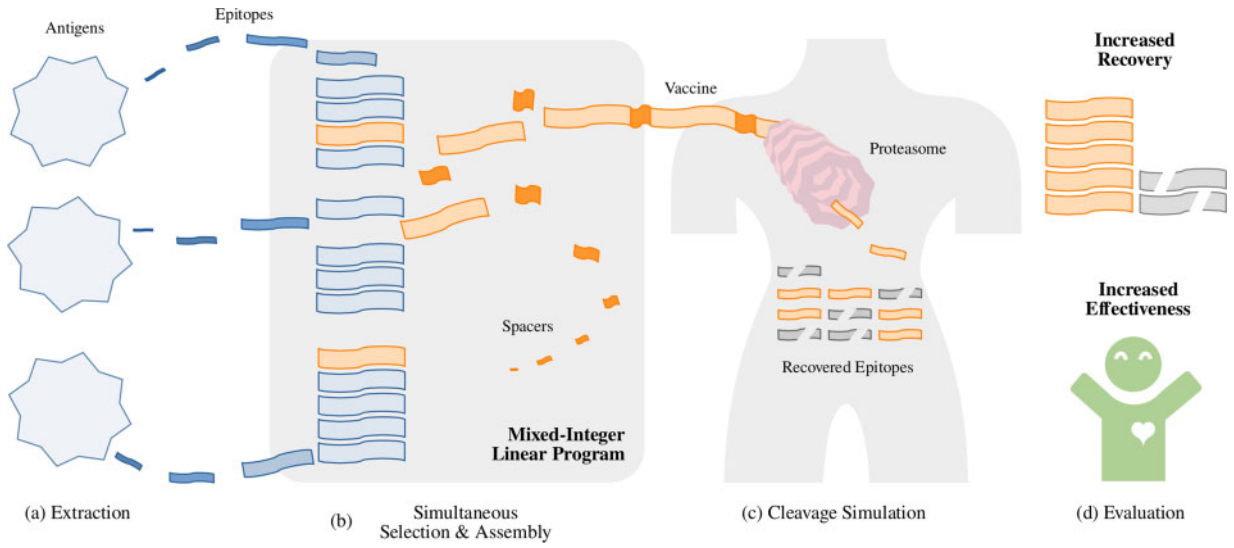

Fig. 1. Conceptual steps in EV design with the proposed framework. (**a**) Epitopes are extracted from a given set of antigens, and properties such as immunogenicity, coverage and conservation are computed. (**b**) We formulate an MILP that creates a string-of-beads vaccine by simultaneously selecting which epitopes to include and how to assemble them into the final construct. This formulation maximizes the immunogenicity of the selected epitopes subject to constraints related to patient and pathogen coverage of the resulting vaccine, as well as cleavage probability of specific residues. To connect the selected epitopes, spacers are designed to provide a high chance of cleavage at the termini of the epitopes, and epitopes that have too high a cleavage probability in their interior are discarded. The vaccine will be subject to proteolytic digestion, which has strong effects on its efficacy. To quantify these effects, (**c**) we perform repeated stochastic simulations of proteasomal cleavage and estimate the probability that each epitope is correctly recovered from the string-of-beads construct. (**d**) Based on the recovered epitopes, the vaccine is evaluated in terms of the average immunogenicity of the recovered epitopes, as well as coverage and conservation with respect to the original antigens and/or the target population. We show that approaching selection and assembly together increases the number of epitopes correctly recovered from the vaccine making the vaccine itself more effective

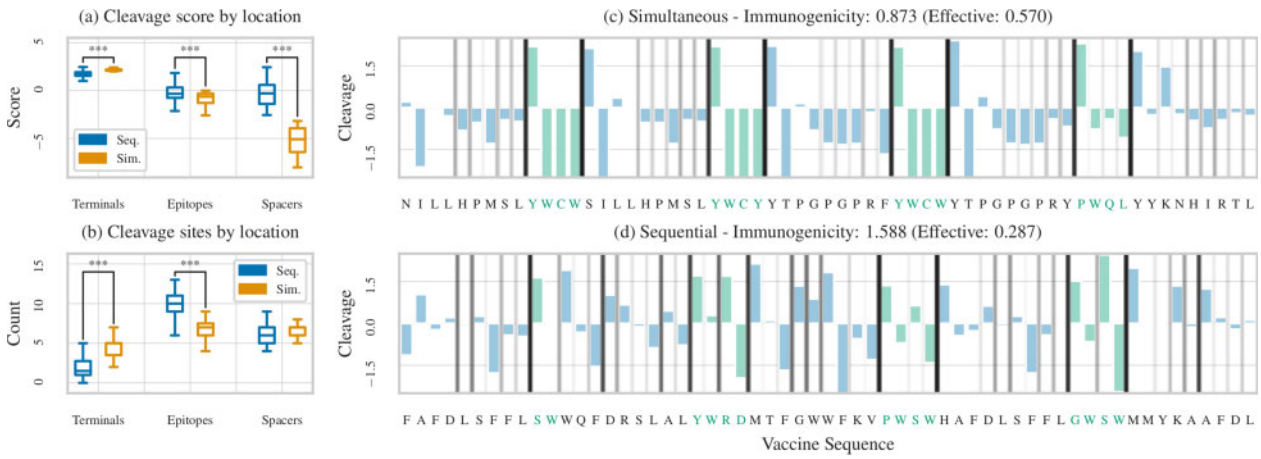

Fig. 2. Comparison of cleavage scores between JessEV and a sequential approach. (**a**) The cleavage scores of residues at the termini, inside the epitopes and inside the spacers for 30 vaccines designed on random subsets of 5000 epitopes. We are able to enforce a strict separation with a clear gap between the scores of residues inside the epitopes and at the termini. (**b**) How many cleavage events, as predicted by NetChop (Nielsen et al., 2005), happened at the termini, inside the epitopes and inside the spacers in the same bootstraps used in (a). The marked differences are statistically significant ($*** < 0.001$). (**c**) The cleavage scores for each residue of a string-of-beads vaccine designed on the complete set of epitopes with a sequential approach and (**d**) with JessEV. The spacers are highlighted in green, and the gray vertical lines represent cleavage frequencies as computed by Monte Carlo simulations with a prior of 0.15, with darker shades being more likely. The title reports both theoretical and effective immunogenicity. Thanks to higher minimum cleavage at the termini and lower maximum cleavage inside the epitopes and spacers, the effective immunogenicity of our vaccine is about twice that of the sequential approach, even though the individual epitopes are less immunogenic

epitope. The bounds we imposed on the cleavage scores forced JessEV to pick epitopes with lower immunogenicity, but thanks to the improved recovery rates the effective immunogenicity was 0.99 ± 0.019 larger than for the sequential approach ($t = 5.240$, $P-\text{value} = 3 \times 10^{-6}$for an effect size of 1.21). However, due to this restriction on the available epitopes, the problem proved infeasible in three cases out of 30. Relaxing the bounds on the cleavage scores would prevent infeasibility in exchange for possibly lower effective immunogenicity.

Comparing vaccines designed with the same procedure on the complete set of epitopes revealed that the epitopes selected by JessEV had a combined immunogenicity of 0.87 (Fig. 2c), only 55% of the immunogenicity of the sequential approach (Fig. 2d). The resulting effective immunogenicity, however, was 0.57, 199% larger than that of the sequential approach. For this and all subsequent experiments, we relaxed the negative cleavage score constraint on the first three residues after the N-terminal of the epitopes. Under our cleavage model, these three residues cannot be cleaved when the N-terminal is, therefore their score is uninfluential as long as the N-terminal is cleaved frequently. This allowed JessEV to pick from a broader variety of epitopes, resulting in a 79% increase in immunogenicity and 64% increase in effective immunogenicity.

3.2 Increased epitope recovery rates improve effective immunogenicity and coverage

We designed string-of-bead vaccines using several thresholds for minimum termini cleavage ν and γ (ranging from 1.5 to 2.5) and maximum epitopes' interior cleavage η (from -1 to 1). To optimize for effective coverage, we additionally performed a grid search on pathogen conservation (from 5% to 20%) while keeping pathogen coverage at 99% and allowing larger values of η (between 1 and 2). We then performed 1000 Monte Carlo cleavage simulations using different prior cleavage probabilities p_c, and selected the values for ν, γ and η that resulted in the largest average effective immunogenicity or coverage for each p_c. Finally, we compared JessEV's best solution with the fixed vaccine produced by the sequential approach (Fig. 3).

The effective immunogenicity of vaccines designed by JessEV was consistently larger by at least 97% than that resulting from a sequential design across all settings of p_c. Often, JessEV's 25th percentile was larger than the 75th percentile of the sequential design (Fig. 3a). The number of recovered epitopes was also at least 315% larger (Fig. 3b). However, the higher epitope recovery rates were partly offset by the lower immunogenicity of the selected epitopes, as forcing low cleavage likelihoods inside the selected epitopes restricted the set of candidates. A qualitatively similar result could be observed for effective pathogen (Fig. 3c) and HLA coverage (Fig. 3d). JessEV's designs outperformed sequential designs by a margin of at least 140% and 27% for effective pathogen and allele coverage, respectively. We were able to design vaccines such that 2.13 epitopes were recovered on average even with $p_c = 1$ by enforcing the interior epitope cleavage score to be smaller than -1, corresponding to a cleavage probability below $e^{-1} \approx 0.37$. In practice, however, half of the residues inside the epitopes of such a vaccine had a cleavage score smaller than -3.91 ($e^{-3.91} \approx 0.02$), showing that favorable cleavage can be achieved even in the most adverse conditions.

We then bootstrapped the Monte Carlo trials to quantify the probability that JessEV's vaccines had worse immunogenicity and HLA/pathogen coverage across different prior recovery probabilities. On average, they were worse 12%, 7% and 1% of the times for effective immunogenicity, pathogen coverage and HLA coverage, respectively (Fig. 3e). Given the low number of HLA alleles, there was a significant probability that both methods could cover the same number of alleles (62% on average and 42% that both effectively covered zero alleles). For pathogens, this probability was lower but still considerable (average 24%), while the effective immunogenicity was equal only 5% of the time.

We also quantified the expected improvement for each p_c (i.e. the ratio between the average effective immunogenicity of the two solutions) and found that JessEV consistently outperformed the state-of-the-art by 2- to 3-fold, with greater improvement as p_c approached one (Fig. 3f).

3.3 The same constraints are effective across a realistic range of prior cleavage probabilities

Different settings of ν, γ and η were needed to obtain the largest possible effective immunogenicity depending on the prior cleavage probability p_c. As p_c increased, ν, γ and η decreased, as the smaller cleavage score was offset by the larger prior probability. Figure 4c traces the evolution of these parameters and shows that for $0.15 \leq p_c < 0.5$ the optimal settings were $\nu = \gamma = 1.95$ and $\eta = -0.1$. A prior between 0.15 and 0.20 resulted in fragments of length between 7 and 10 on our dataset, consistent with what was observed *in vitro* (Nussbaum *et al.*, 1998), suggesting that these are promising values for experimental testing. Additionally, Figure 4a shows that for these prior probabilities the effective immunogenicity of the second,

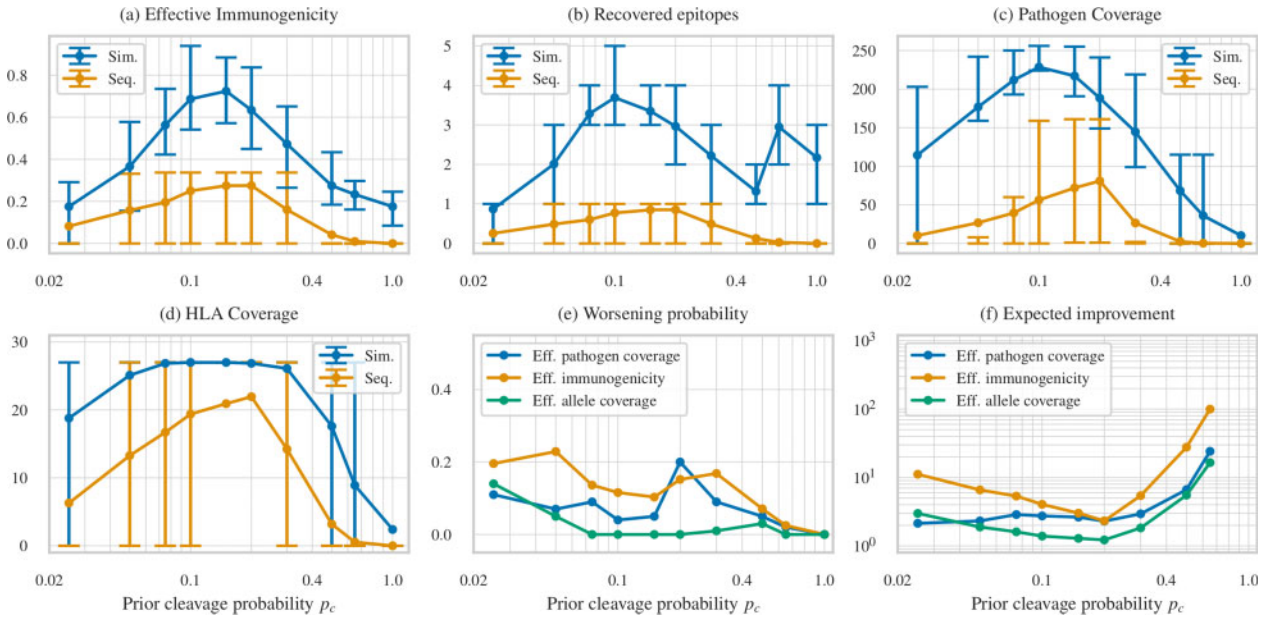

Fig. 3. Evaluation of string-of-beads designed with JessEV and a sequential approach across different prior cleavage probabilities. (**a**), (**b**), (**c**) and (**d**) Mean, 25th and 75th percentile of the Monte Carlo simulations for effective immunogenicity, recovered epitopes, pathogen coverage and HLA coverage, respectively. JessEV was better under all metrics across all choices of prior cleavage probabilities. Note that both vaccines optimized for immunogenicity in (a) and (b), and for coverage in (c) and (d), which means that different constraints were used to produce them. (**e**) and (**f**) The probability of worsening and expected improvement of effective immunogenicity, effective pathogen coverage and effective HLA coverage. Both were estimated through 5000 bootstrap of the outcomes of the 1000 Monte Carlo simulations. String-of-bead vaccines produced by JessEV were very frequently not worse than the sequential approach, and on average between three to five times better across a realistic range of prior probabilities. At cleavage probabilities larger than 0.7, no epitopes were ever recovered for the sequential approach; hence, the expected improvement approached infinity

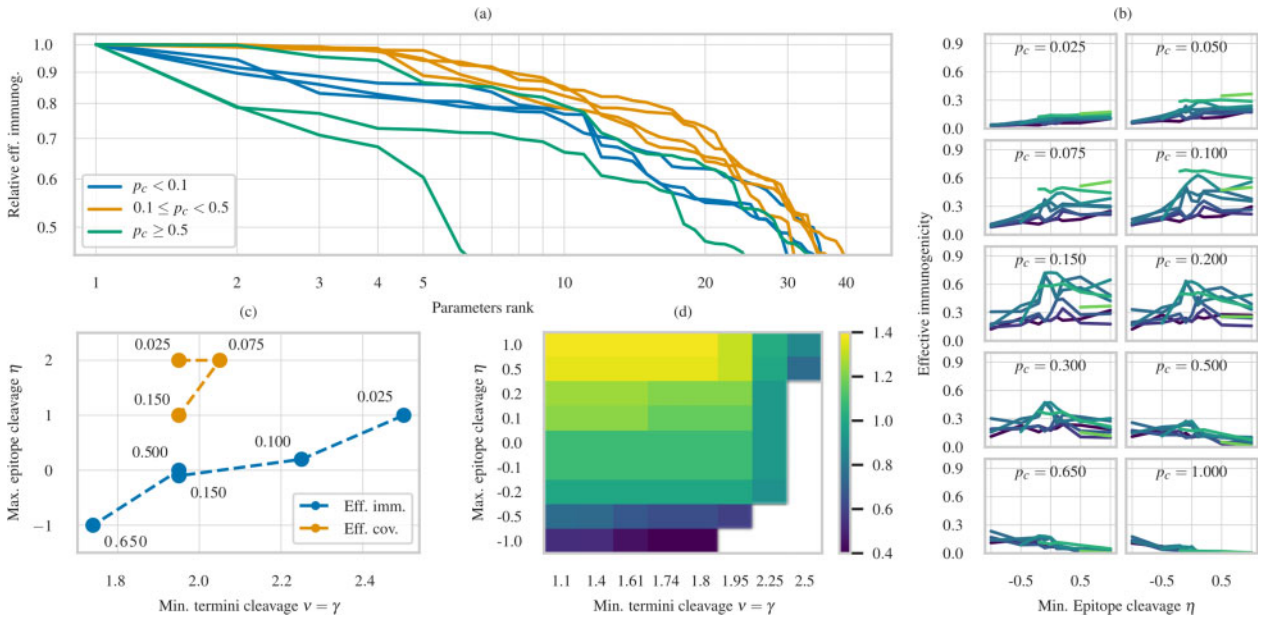

Fig. 4. The effects of cleavage constraints on the immunogenicity objective and effective immunogenicity. (**a**) For each prior probability, we show the effective immunogenicity relative to the best obtained for that prior (*y*-axis) for different parameter settings ranked by effective immunogenicity (*x*-axis). There is a range of prior probabilities, from 0.1 to 0.3, where four or five different parameter settings were within 5% of the largest effective immunogenicity. (**b**) Effective immunogenicity (*y*-axis) as a function of the inner epitope cleavage (*x*-axis) for different cleavage at the termini (lighter lines for larger constraints) and nine different prior cleavage probabilities (in each sub-figure). For prior cleavage probabilities in a reasonable range, the best effective immunogenicity was obtained with an inner epitope cleavage around zero, while lower settings worked best for high priors and larger ones for low priors. (**c**) The effect of prior cleavage probability (annotated close to each data point) on parameters (*x*- and *y*-axes) that resulted in the largest effective immunogenicity (blue) or effective pathogen coverage (orange). Only transitions are displayed, meaning that several prior probabilities between, e.g. 0.15 and 0.5 (not included) had the same optimal settings for the effective immunogenicity. As the prior cleavage probability increased, constraints on cleavage at the termini could be relaxed, while the score inside the epitopes must be kept lower. Optimizing for effective coverage required larger possible cleavage likelihoods inside the epitopes, but similar cleavage likelihoods at the termini. (**d**) Immunogenicity objective for different cleavage constraints, with light background for infeasible settings. Enforcing low cleavage likelihoods inside the epitopes greatly reduced the immunogenicity objective, as many epitopes are not eligible due to higher cleavage likelihoods in the residues of their second half, which cannot be reduced through the preceding spacer under our cleavage model

third and fourth best settings were within 5% of the best design for that prior, meaning that the effective immunogenicity for these priors was not overly sensitive to the settings of ν, γ and η.

In general, η was more critical than ν and γ, since it had a considerable effect on the set of epitopes that could be selected (Fig. 4d). In fact, according to our cleavage model, a spacer affects cleavage only in the first four residues of the following epitope, while the score in the following five residues cannot be altered. For prior cleavage probabilities in a reasonable range, the largest effective immunogenicity was obtained with η close to zero, whereas larger absolute values caused a reduction in the effective immunogenicity. For very high (low) prior probabilities, the best results were obtained with low (high) values for η (Fig. 4b).

This also explains why optimizing for effective coverage required larger values for η: as very few epitopes were conserved across a sufficient number of pathogens, they were excluded when η was too small. In fact, the epitopes with the highest pathogenic coverage raking in the top 1%, 2% and 5% percentile covered 21%, 13% and 5% of the pathogenic antigens, respectively. This illustrates that including conserved epitopes is fundamental even if they are recovered less frequently.

3.4 Optimized spacers are necessary but few variants are used

Inspecting the vaccines with largest effective immunogenicity for each of the 10 prior cleavage probabilities revealed that only nine different spacer sequences were used. Two sequences, MWQW and MWRW, were used in 19 out of 40 spacers. These spacer sequences increase the C-terminal cleavage score by 2.37 and 2.32, respectively, corresponding to a 10-fold likelihood increase. Only five of the possible 20^4 sequences induce a larger increase. However, they all end in K, which reduces the N-terminal score by 1.4 and greatly limits the number of viable epitopes after the spacer.

Designing vaccines with fixed MWQW spacer resulted in a reduction in immunogenicity, both theoretical and effective, of 35–45% compared to the optimized spacers. Similarly, using the popular spacer AAY caused a decrease of 85–95%, and required to relax the cleavage constraints on ν and γ from 1.95 to 1.6 and η from 0 to 1.0 for the problem to be feasible at all. This highlights the need for spacers designed *ad hoc*.

3.5 Runtime analysis

The linear programs were solved on a cloud virtual machine limited to 10 out of 20 cores of an Intel Xeon Gold 6148 CPU. Half of the experiments completed within 9 min, and 90% within 56 min. The slowest 1% of the experiments required 5 h or more, and 83% (58%) of them spent more than 50% (75%) of the time closing a 10% gap between the lower and upper bounds for the optimal cost.

4 Discussion and conclusions

No current state-of-the-art design approach is able to simultaneously select epitopes and assemble them into a string-of-beads vaccine construct. Our work fills this gap through a linear programming formulation that guarantees optimality of the design. This linear program finds a set of epitopes of maximal immunogenicity, as well as their arrangement and spacers linking them, ensuring that the vaccine satisfies constraints related to pathogen, HLA coverage, conservation and cleavage likelihood in critical positions of the construct.

Being based on MILP renders, the simultaneous epitope selection and assembly problem of string-of-beads vaccines NP-hard. In most cases, this does not prevent the solver to find a solution in $<$1 h, in part because of the many heuristics (Berthold, 2006; Bixby *et al.*, 2000; Fischetti and Lodi, 2011) that can be employed. Though, certain constraint configurations can make the solving process much slower, up to 5 h. In these rare cases half or more of the time is usually spent on improving a solution whose objective value is already within 5 to 10 percentage points to the optimum. As this gap is known during the solving process, the solver can be interrupted early, obtaining an almost-optimal solution with formal guarantees on its quality. Indexing the PSSM to compute cleavage scores contributes a great deal to the overall complexity of the linear program, as its size in terms of variables and constraints grows quadratically with the maximum number of residues in the vaccine. Using spacers of fixed length can therefore significantly reduce the computational resources needed to find a vaccine, at the price of possibly longer and less well-cleaved polypeptides.

We assumed a simple stochastic model of proteasomal cleavage and used Monte Carlo simulations to estimate the recovery probability of each epitope in a vaccine to show that approaching the epitope selection and epitope assembly problems together resulted in increased recovery probability of the epitopes in the vaccine. We also verified our results with NetChop Cterm (Nielsen *et al.*, 2005), an independent proteasomal cleavage prediction tool, to confirm that, in spite of the simplistic nature of the cleavage predictor we used in the linear program, our approach significantly reduced the number of cleavage sites inside the epitopes and increased the cleavage frequency at their termini. This, in turn, translated to improved effective immunogenicity, coverage and conservation. The main reason for this improvement can be traced to the ability of our framework to select epitopes that have a small cleavage probability in their interior, thus preventing unwanted cleavage in these locations. We also argued that this constraint should be relaxed in order to include highly conserved epitopes, as the gains in coverage offset the reduction in recovery frequency.

Our framework depends on the ability to express the computation of the cleavage score in a linear form. Non-linear functions can be approximated by piece-wise linear approximations, and lookup tables can be used in the worst case, but this could make solving even small instances of the linear program impractical due to extremely long runtimes. As a precaution against the possibility that our cleavage model is too simplistic, stricter cleavage constraints than what our simulations suggested could be enforced. Moreover, finding the right bounds for the cleavage scores requires an outer optimization loop where the vaccines are evaluated using the Monte Carlo simulations. In this work, we performed a grid search to study the influence of the parameters on the effective immunogenicity, but more complicated optimization strategies such as Bayesian optimization (Brochu *et al.*, 2010; Shahriari *et al.*, 2016) can be used to reduce the computational requirements needed to find solutions with good effective immunogenicity or coverage.

In conclusion, our approach allows precise control of the cleavage probability of every residue in a string-of-beads construct through simultaneously approaching epitope selection and vaccine assembly. This enables significant improvements in the recovery probability of the epitopes in the construct, which translates to increased effectiveness of the vaccine as a whole.

Funding

E.D. was supported by the Helmholtz Association under the joint research school 'Munich School for Data Science' (MuDS, Award Number HIDSS-0006 from the Helmholtz Association). B.S. acknowledges financial support by the Postdoctoral Fellowship Program of the Helmholtz Zentrum München.

Conflict of Interest: none declared.

References

Audran,R. *et al.* (2005) Phase I malaria vaccine trial with a long synthetic peptide derived from the merozoite surface protein 3 antigen. *Infect. Immun.*, **73**, 8017–8026.

Barouch,D.H. *et al.* (2018) Evaluation of a mosaic HIV-1 vaccine in a randomized, double-blinded, placebo-controlled phase I/IIa clinical trial and in rhesus monkeys. *Lancet*, **392**, 232–243.

Berthold,T. (2006) Primal heuristics for mixed integer programs. Master's Thesis, Zuse Institute Berlin, Berlin, Germany.

Bixby, R.E., *et al.* (2000) MIP: Theory and Practice — Closing the Gap. In: Powell M.J.D., Scholtes S. (eds) *System Modelling and Optimization. CSMO* **1999**. IFIP — The International Federation for Information Processing, vol 46. Springer, Boston, MA.

Brochu,E. *et al.* (2010) A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*

Cornet,S. *et al.* (2006) Optimal organization of a polypeptide-based candidate cancer vaccine composed of cryptic tumor peptides with enhanced immunogenicity. *Vaccine*, **24**, 2102–2109.

Dönnes,P. and Kohlbacher,O. (2005) Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci.*, **14**, 2132–2140.

Fischetti,M. and Lodi,A. (2011) *Heuristics in Mixed Integer Programming*. Wiley Encyclopedia of Operations Research and Management Science, John Wiley & Sons, Inc.

Foley,B.T. *et al.* (2018) HIV Sequence Compendium 2018. *Technical Report LA-UR-18-25673*. Los Alamos National Lab. (LANL), Los Alamos, NM, USA.

Gurobi Optimization,L. (2020) Gurobi optimizer reference manual. http://www.gurobi.com.

Hart,W.E. *et al.* (2011) Pyomo: modeling and solving mathematical programs in python. *Math. Program. Comput.*, **3**, 219–260. Springer Science and Business Media LLC.

Hart,W.E. *et al.* (2017) *Pyomo–Optimization Modeling in Python*. Vol. **67**, 2nd edn. Springer Science & Business Media.

Hu,Z. *et al.* (2018) Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat. Rev. Immunol.*, **18**, 168–182.

Hunter,J.D. (2007) Matplotlib: a 2d graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.

Jurtz,V. *et al.* (2017) NetMHCpan-4.0: improved peptide–MHC class i interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.*, **199**, 3360–3368.

Kuttler,C. *et al.* (2000) An algorithm for the prediction of proteasomal cleavages. *J. Mol. Biol.*, **298**, 417–429.

Li,B.-Q. *et al.* (2012) Prediction of protein cleavage site with feature selection by random forest. *PLoS One*, **7**, e45854.

Liu,M.A. (2019) A comparison of plasmid DNA and mRNA as vaccine technologies. *Vaccines*, **7**, 37.

Los Alamos National Laboratory (2019) The HIV sequence database (03 October 2019, date last accessed).

Lundegaard,C. *et al.* (2010) PopCover: a method for selecting of peptides with optimal population and pathogen coverage. In: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology—BCB'10, ACM Press, Niagara Falls, NY, USA, p. 658.

McKinney,W. (2010) Data structures for statistical computing in python. In: van der Walt, S. and Millman, J. (eds.) *Proceedings of the 9th Python in Science Conference*, pp. 51–56.

Nielsen,M. *et al.* (2005) The role of the proteasome in generating cytotoxic t-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, **57**, 33–41.

Nussbaum,A.K. *et al.* (1998) Cleavage motifs of the yeast 20S proteasome β subunits deduced from digests of enolase 1. *Proc. Natl. Acad. Sci. USA*, **95**, 12504–12509.

Ott,P.A. *et al.* (2017) An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, **547**, 217–221.

Paul,S. *et al.* (2013) HLA class I alleles are associated with peptide binding repertoires of different size, affinity and immunogenicity. *J. Immunol.*, **191**, 5831–5839.

Pérez,F. and Granger,B.E. (2007) IPython: a system for interactive scientific computing. *Comput. Sci. Eng.*, **9**, 21–29.

Peters,B. *et al.* (2020) T cell epitope predictions. *Annu. Rev. Immunol.*, **38**, 123–145.

Sahin,U. and Türeci,Ö. (2018) Personalized vaccines for cancer immunotherapy. *Science*, **359**, 1355–1360.

Schubert,B. and Kohlbacher,O. (2016) Designing string-of-beads vaccines with optimal spacers. *Genome Med.*, **8**, 9.

Schubert,B. *et al.* (2016) FRED 2: an immunoinformatics framework for python. *Bioinformatics*, **32**, 2044–2046.

Seabold,S. and Perktold,J. (2010) statsmodels: econometric and statistical modeling with python. In: 9th Python in Science Conference.

Sette,A. *et al.* (1994) The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J. Immunol.*, **153**, 5586–5592.

Shahriari,B. *et al.* (2016) Taking the human out of the loop: a review of Bayesian optimization. *Proc. IEEE*, **104**, 148–175.

Tenzer,S. *et al.* (2005) Modeling the MHC class i pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class i binding. *CMLS Cell. Mol. Life Sci.*, **62**, 1025–1037.

Toussaint,N.C. *et al.* (2008) A mathematical framework for the selection of an optimal set of peptides for epitope-based vaccines. *PLoS Comput. Biol.*, **4**, e1000246.

Toussaint,N.C. *et al.* (2011) Universal peptide vaccines—optimal peptide vaccine design based on viral sequence conservation. *Vaccine*, **29**, 8745–8753.

van der Walt,S. *et al.* (2011) The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.*, **13**, 22–30.

van Rossum,G. (2001) Python reference manual. *Technical Report*, PythonLabs, Virginia, USA.

Velders,M.P. *et al.* (2001) Defined flanking spacers and enhanced proteolysis is essential for eradication of established tumors by an epitope string DNA vaccine. *J. Immunol.*, **166**, 5366–5373.

Vielma,J.P. *et al.* (2010) Mixed-integer models for nonseparable piecewise-linear optimization: unifying framework and extensions. *Oper. Res.*, **58**, 303–315.

Virtanen,P. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods*, **17**, 261–272.

Waskom,M. *et al.* (2017) Mwaskom/seaborn: V0.8.1 (September 2017) https://zenodo.org/record/54844 (September 2020, date last accessed).

Yang,B. *et al.* (1996) The requirement for proteasome activity class I major histocompatibility complex antigen presentation is dictated by the length of preprocessed antigen. *J. Exp. Med.*, **183**, 1545–1552.