Computational Antigen Optimization through Symbolic Optimization and Affinity Maturation Simulation

Jonathan G. Faris[1,2], **Mikel Landajuela**[1], **Kayla G. Sprenger**[2], **Daniel Faissol**[1], **and Felipe Leno da Silva**[1]
[1] Lawrence Livermore National Laboratory, Livermore, USA.
[2] University of Colorado, Boulder, USA.
{jonathan.faris,kayla.sprenger}@colorado.edu
{landajuelala1, faissol1, leno}@llnl.gov

Abstract

With the recent, significant improvement of computational tools for protein interaction prediction, the use of machine learning to support the development of vaccination regimens brings with it new hope for diseases which, so far, have eluded our best efforts at finding a cure, like HIV. We here propose BIOVAX, a novel pipeline combining symbolic optimization with affinity maturation simulation to generate highly-optimized antigens intended for vaccination development. We perform an *in silico* evaluation using real HIV targets, and show that the antigen designed by BIOVAX elicit estimated antibodies that bind more strongly to a diverse, global panel of real HIV viruses than both the parent sequence, and other computationally-designed antigen baselines available in the literature. BIOVAX is our first step towards a new generation of AI-assisted vaccine development pipelines.

1 Introduction

The development of an effective vaccine for human immunodeficiency virus-1 (HIV) has remained elusive, primarily due to the virus' ability to evade the adaptive immune response by rapidly evolving into a diverse viral population within the host. Over the past few decades, the isolation and characterization of broadly neutralizing antibodies (bnAbs) from HIV-infected individuals have generated optimism that the immune system may indeed be capable of defending against this challenging infection. bnAbs are remarkable within the human antibody repertoire for their ability to retain neutralization potency across a wide variety of pathogen strains, even as viral surface proteins (antigens) mutate. This unique capability has sparked hope that these antibodies could be induced via vaccination, providing robust protection against HIV. Since their initial discovery, a plethora of bnAbs have been identified [22]. However, despite their success in controlling viral populations within their natural hosts, the induction of bnAbs through vaccination has yet to be demonstrated *in vivo* [23].

Traditional approaches to antigen design have relied on empirical evidence of neutralizing sera in response to live-attenuated or inactivated viruses [39]. With the rise in availability of high-quality structural information, we have entered a new era of structure-based vaccine design, which fueled the rapid development of vaccines against SARS-CoV-2 in response to the COVID-19 pandemic [33].

Unfortunately, pathogens such as HIV presents additional challenges that have yet to be fully addressed, including substantial antigenic variability, genomic flexibility, and extensive, diverse glycosylation patterns. The high degree of immune evasion or escape suggests that a vaccine must

AI for New Drug Modalities at NeurIPS 2024.

either induce especially broad bnAbs or a diverse, potent polyclonal response to be sufficiently protective against HIV.

Given the lack of success with traditional vaccine design approaches [23] and the rapidly growing suite of bioinformatics software distributions [5, 9, 20, 47], computationally-based approaches are on the rise. Building upon foundational work on stabilizing the gp160 trimer [27, 41] and previous efforts in developing mathematical models of affinity maturation [12, 17, 48]–the Darwinian process of *in vivo* antibody evolution–we propose a novel pipeline for antigen optimization. This pipeline features a machine learning (ML)-based antigen sequence optimizer allied with simulations of affinity maturation to downselect promising vaccine candidate antigens. We primarily focus on HIV as a model pathogen, but the pipeline is general.

In Section 2, we first review relevant background information on vaccine design, affinity maturation, and computational protein redesign. We then formalize the problem and approach addressed in this study in Section 3. In Section 4, we describe in details our proposed pipeline, while in Section 5 we present the empirical evaluation of our proposed pipeline using real HIV targets. Finally, we conclude the paper in Section 6.

2 Background

2.1 Vaccine Design and Development

Vaccination has proven to be one of the most successful approaches to disease prevention, control, and treatment to date [31]. Vaccines typically contain either a live-attenuated virus, viral vectors, protein subunits, or, more recently, mRNA encoding a pathogenic protein [19, 38]. Each of these approaches comes with potential benefits and pitfalls that must be considered in the design process. Live-attenuated vaccines, such as the Polio vaccine, have been shown to induce a robust immune responses, resulting in potent neutralization in the resulting sera, but they may also carry a risk of breakthrough infections. Most notably, while almost completely eradicated from the human population, cases of Polio still persist–not due to natural infection, but because of breakthrough infections following vaccination[1].

Protein-subunit immunogens mitigate the risk of breakthrough infections, as they contain only one or a few antigens, rather than the full set of viral machinery. However, they can suffer from reduced immunogenicity, necessitating multiple vaccinations to confer protection, similar to what was seen with the prime-boost strategy used against SARS-CoV-2 [3, 37]. Additionally, traditional protein-subunit vaccinations represent only a single time point in the evolutionary trajectory of a highly mutable pathogen. These pathogens may therefore evolve to evade the immune response with time, requiring updates to the antigen sequence and additional immunizations at regular intervals (e.g., the annual influenza vaccine [40]).

Finally, mRNA technologies offer the advantage of rapid development and strong immunogenicity by directly delivering the genetic instructions for the host cells to produce the target protein, which ensures proper folding and post-translational modifications within the host, as well as continued expression of the antigen [38]. However, mRNA vaccines may present challenges related to the stability of the mRNA itself, which requires cold-chain logistics and careful formulation to prevent degradation before delivery [34]. This need for stabilization adds a critical bottleneck to the development and distribution of mRNA vaccines [29].

While all of these strategies are viable, each may be better suited to particular pathogens or diseases. For example, the risk of breakthrough infection associated with live-attenuated viruses may be too concerning for HIV, but less so for influenza. Given our current focus on designing antigens for the HIV surface receptor-binding protein, glycoprotein 120 (gp120) [41], we elected to utilize a protein-subunit vaccine design approach. This choice not only avoids the potential risk of breakthrough infections but also avoids the complexity associated with stabilizing and expressing mRNA constructs. By focusing on a gp120 subunit antigen, we can leverage tractable computational models of the adaptive immune response for evaluation and screening in the proposed design pipeline.

[1] https://www.cdc.gov/vaccines/vpd/polio/hcp/vaccine-derived-poliovirus-faq.html

2.2 Affinity Maturation

Affinity maturation (AM) is the evolutionary process by which antibodies evolve in response to natural infection or vaccination [43, 50]. Following exposure, B cells with receptors (BCRs) that have low affinity for the antigen seed microstructures within lymph nodes known as germinal centers (GCs). GCs can be spatially divided into two regions: the dark zone and the light zone. In the dark zone, B cells proliferate, and their BCRs undergo mutations induced by activation-induced cytosine deaminase (AID) in a process known as somatic hypermutation (SHM). After SHM, B cells migrate to the light zone, where follicular dendritic cells (FDCs) display the antigen. B cells with functional BCRs then participate in an affinity-based competition to bind the antigen, pry it from FDCs, and internalize the resulting immune complex. After successfully internalizing and breaking down the antigen, B cells present the resultant antigen peptides on their surfaces, which are recognized by helper T (T_H) cells. T_H cells then signal the B cells to either (1) return to the dark zone; (2) differentiate into memory B cells to protect against future exposures; or (3) exit the GC and enter the plasma to fight the infection by producing soluble BCRs (antibodies; Abs). Upon receiving survival signals from T_H cells, most B cells recycle back to the dark zone to continue accruing mutations.

When exposed to a single antigen, AM has been shown to result in predominantly "strain-specific" antibodies (i.e., antibodies capable of neutralizing a narrow subset of viral strains closely related to the antigen administered in the vaccine or initially encountered during natural exposure) [1]. Conversely, delivering sequential immunizations with optimally and increasingly variant antigens is expected to focus BCR evolution on regions of the antigens that are conserved across multiple strains [12], yielding broadly neutralizing antibodies (bnAbs), which are unique in their ability to recognize a diverse array of viral strains.

The development of an effective vaccine requires numerous labor-intensive steps including, but not limited to: (1) *in vitro* expression, purification, and stabilization of the antigen; (2) *in vivo* studies of animal models characterizing the adaptive immune response, and the level of protection conferred in response to infection challenges; (3) studies in non-human primates (NHP); and (4) clinical trials for both safety and efficacy in humans.

Each one of these steps adds their own unique challenges, difficulties, and expenses. To increase the likelihood of translational success, computational pipelines offer a safe and cheap way to screen vaccine candidates prior to wet-lab experiments.

2.3 Computational Redesign of Proteins

Protein design and engineering has, until recently, relied primarily on volume-based methods such as directed evolution where large mutant libraries are generated, screened, and regenerated until a sufficiently improved protein has been developed [2, 8]. Directed evolution has been wildly successful in aiding the design of proteins with a plethora of properties–from enzyme activity to fluorescence [15, 58]. Rational design, on the other hand, involves using minimal wet-lab assays, and instead relies on structural information and expert knowledge to engineer improved desired properties [30]. Both approaches, however, still require labor-intensive experiments rendering designing novel protein therapies and enzymes inefficient and costly.

The recent increases in bio-process modelling efficiency has enabled the development of *in silico* protein design platforms [4, 6, 18, 28, 46, 52]. Notably, AlphaFold has revolutionized our ability to make structurally informed insights of proteins on the nanoscale. More recently, some works have demonstrated the translational capability of computational protein-design pipelines [14]–greatly accelerating the speed with which proteins may be rationally designed and screened [28].

Recently, Deep Symbolic Optimization (DSO) [35], has been used with success as, amongst other applications [36, 44], an antibody redesigner [46, 45] typically by improving an existing antibody towards improved binding to a pathogen of interest. DSO models protein redesign as a Symbolic Optimization task, as explained in the remainder of this section. We adapted DSO to redesign antigens rather than antibodies, as further detailed in Section 4.

DSO searches solutions consisting of a discrete, symbolic sequence of tokens to maximize a scoring function. Starting from a library of tokens, $\mathcal{L} = \{\lambda^1, \ldots, \lambda^n\}$, a sequence, τ, can be built where $\tau = \langle \tau_1, \ldots, \tau_n \rangle$ (with τ_i representing the token at position i). The sequence represents a potential

solution to the problem under investigation. In general, τ may be of any length, and may contain copies of the same token, λ^i. After generating a potential solution sequence, a scoring function, or reward signal, is then computed $\mathcal{R} : \tau \rightarrow \mathbb{R}$. All token combinations resulting in valid sequences can then be scored according to their fitness via the reward function $\mathcal{R}(\tau)$. Thus, the general solution to a symbolic optimization problem takes the form:

$$arg\,max_{n\in\mathbb{N},\tau}[\mathcal{R}(\tau)] \text{ with } \tau = \langle \tau_1, \ldots, \tau_n \rangle, \text{ and where } \tau_i \in \mathcal{L} \tag{1}$$

That is, DSO attempts to uncover the sequence which optimizes the reward function. This is typically performed by optimizing the Risk-Seeking Policy Gradient cost:

$$J(\theta) := \mathbb{E}_\theta \left[R(\tau) \mid R(\tau) \geq Q_\epsilon \right], \tag{2}$$

Therefore, DSO aims at searching a huge space of token sequences efficiently. For antibody design specifically, each token τ consists of an amino acid (from the set of 20 natural amino acids), which concatenated form the whole antibody sequence to be sampled by DSO. The reward function is typically an estimation of the binding strength to the pathogen to be targeted (although some exploratory works have considered additional objectives such as stability and humanness [18]).

Therefore, DSO samples protein aminoacid sequences and learns how to associate sequences to their binding strength to a particular target, which leads us to improved versions of existing antibodies.

3 Problem Description

Isolates obtained from patients [53, 59] have demonstrated our natural ability to generate potent bnAbs against HIV. Therefore, it is reasonable to assume that for many diseases we might already have identified a bnAb we want to elicit in the general population upon vaccination. Likewise, convergent BCR gene usage has been observed for a variety of diseases in the population, suggesting commonalities in the human GL repertoire [25].Consequently, we assume we also have identified the unmutated common ancestor (UCA) of our target antibody (which can be done via tools such as IgBlast [56]). We then assume this will be the starting point of the immune system for the simulated vaccination. However, developing an antigen capable of eliciting a desirable immune response is still a very difficult problem. By starting from a known, patient-derived bnAb, we are able to search the literature to identify an already-existing promising antigen, which will be considered by us our starting point for the designed antigens[2].

We also assume that the mutation and substitution probability distributions of BCRs are known, so that we can estimate the mutations that would be introduced *in vivo*. Those probabilities have already been characterized in the literature [55].

We also assume we have a way of quickly estimating the free-energy of binding (ΔG) between arbitrary antigens and antibodies. There is a myriad of simulation tools able to perform this task with varied levels of precision and costs [6, 10, 24, 54].

Given the inputs described above (an identified bnAb, the corresponding GL antibody sequence, an initial antigen sequence, and a binding estimation tool) the problem we are trying to solve in this paper can be described as follows: find a modified version of the antigen in a way that we increase the probability of a vaccinated person producing a bnAb (be it the exact same sequence or another with comparable or better binding capabilities).

All of this process is to be performed *in silico*, while taking into account realistic computational budgets. We believe this approach will provide novel insights into both the underlying immune dynamics in response to complex antigens, as well as being realistic a proof-of-concept for a novel vaccine design pipeline.

4 Vaccine Development through Antigen Redesign

We describe our approach in details in this section. To aid in the rapid, cost-effective development of novel vaccine antigens against highly mutable pathogens (i.e., solving the problem described in the

[2]This starting antigen is assumed to have additional desirable properties for a vaccine antigen, such as stability, solubility, manufacturability, and so forth.

last section), we propose to tackle this problem with our design pipeline, BIOVAX (**B**ioinformatics-**O**ptimized **V**accine **A**ntigen e**X**plorer).

BIOVAX is illustrated in Figure 1 and works as follows. Starting from the known antigen-antibody binding pair, and the corresponding germline, a protein-sequence optimizer will propose a number of modified antigen designs by introducing mutations in the original antigen. The primary metric followed by the optimizer is the binding affinity between the design and the GL antibody. We hypothesize that optimizing for tighter binders to the GL will increase the likelihood rare bnAb precursors will be activated. After a diverse set of antigen designs is generated by the optimizer, the final screening will be performed by the affinity maturation (AM) module.

The AM model will then predict the resulting panel of antibodies a person will develop when vaccinated with each one of the designed antigens (single vaccination). In the end, antigens are selected by providing better HIV protection according to the AM simulation via their predicted improvements in breadth.

Using the workflow outlined above, we hope this pipeline will alleviate some of the expensive, cumbersome, and empirical process of vaccine optimization in a manner which compliments and accelerates the *in vitro* and *in vivo* efforts in this field. In the following subsections, we discuss in more detail the steps in our workflow, the assumptions underlying the model, and the rationale behind critical components of the maturation simulation.

4.1 Design of antigens using Deep Symbolic Optimization

Our proposed antigen optimizer is an adapted version of the Deep Symbolic Optimization (DSO) platform (Section 2.3), where we model our problem similarly as it was reported to be used in the engineering of anti-SARS-CoV2 antibodies [45]. Our motivation of using DSO stemmed from the knowledge that our antigen optimization step is in many ways similar to antibody optimization, a task where DSO was shown the excel. Instead of mutating an initial antibody and producing candidate antibody designs, we design novel antigen candidates by mutating our initial antigen.

To define the antigen residues that can be mutated by DSO we determine a contact surface between the antigen and the GL antibody (using a distance cutoff) from its structure (which either is available

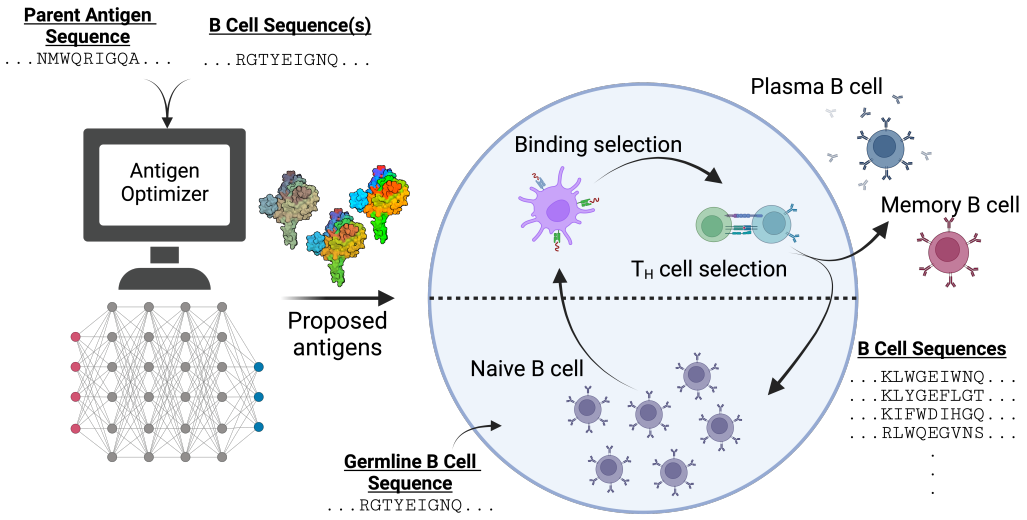

Figure 1: Overview of the proposed antigen design pipeline, BIOVAX. Primary inputs and outputs of the process are bolded and underlined. Given an initial broadly-neutralizing antibody, its corresponding GL antibody, and an initial Antigen, the *antigen optimizer* will design a set of proposed antigens that optimize binding to the GL Ab. Those antigens will be evaluated and ranked in an affinity maturation simulation, and the antigen with better estimated resulting protection will be the output of the pipeline.

from a solved crystal structure or can be estimated through the use of tools such as AlphaFold2 [28] or RoseTTAFold [4]).

The remainder of the optimization strategy is executed, as reported in the literature [18, 35, 46, 45], by optimizing the sampled antigens according to estimated binding affinity between the GL antibody and the candidate antigen.

4.2 Modeling of Affinity Maturation

The affinity maturation model used herein has been described in rigorous detail elsewhere [12, 48], here we aim to provide a brief overview of what the model inputs are, and what the model is broadly doing. Briefly, the primary inputs to the model are: (1) the nucleotide sequence of the relevant GL BCR; (2) the amino acid sequence for the proposed antigen; (3) the number of antigen copies, representing a pseudo-concentration; (4) The structure of the complex and, if applicable, additional information required by the binding simulation (5) probability tables for (a) making a mutation at a given position, and (b) likely substitutions at this position. The simulation is then initialized by generating GCs which contain B cells with the relevant GL sequence(s). B cells then undergo proliferation and somatic hypermutation, simulating the GC dark zone replication and activation of AID [42]. After acquiring mutations, the ΔG of each BCR population is computed, and the binding affinity is then used to determine the amount of antigen each B cell is able to internalize (see below for details). Following binding selection, help from T_H cells is sought and allocated. The B cells then can recycle for further rounds of maturation, or exit the GC and differentiate into (a) plasma B cells, or (b) memory B cells to be activated upon subsequent immunizations. The remainder of this section will outline additional important details of the model, followed by system-specific parameters and experimental procedures in Section 5.

Germinal centers are initially seeded with a single germline (GL) B cell lineage from the identified mature bnAb family targeting. The GL population begins with 100 cells split evenly across ten identical clonal populations. Additionally, to model vaccination or exposure, the Ag amino acid sequence is provided when seeding the GC to stimulate the immune response. While the selection criteria operates on the level of the protein amino acid sequence, mutations are introduced via AID into the BCR nucleotide sequence in a biased manner [55].

To gain insights into possible paths the antibody sequence may take, we provide the model with the nucleotide sequence of the naive BCR. Recently, Yaari et al. developed an emprical model characterizing the relative probability of mutation (mutability) at a given nucleotide position, and the likelihood of each nucleotide substitution at the central location given the local sequence identity [55]. By examining over 1 million sequences before and after AM, they determined AID depends on the local nucleotide environment in five nucleotide segments (fivemers). To capture these dynamics, we determine all relevant fivemers via a sliding window then obtain the corresponding mutability via a simple lookup table. The mutability and substitution scores are then normalized, and a mutation is selected based on these probabilities via a random binomial.

Mutations are introduced at a rate of 0.14 mutations per sequence per division [7]. Further, to reduce the computational load at runtime we consider only mutations in the complementarity-determining regions (CDRs) are considered to impact the BCR-Ag binding free-energy (ΔG). For the purposes of this study, framework region (FRW) mutations were not modeled, as our scoring function does not capture potential changes in stability or flexibility associated with these mutations [32]. We set the probability of mutating the CDR, P_{CDR}, to 0.85, and assume 30% of all mutations to be lethal [57].

Upon generating mutant populations in each GC cycle, B cells with functional BCRs then compete with one another for the binding and internalization of Ag before seeking aid from T_H cells. The level of internalization has been shown to influence the number of times a particular clone will divide to generate progeny [21]. To represent the relationship between binding affinity, available Ag, and the probability of internalization we propose the following:

$$P_{capture} = (l - \frac{Ag_i}{1 + Ag_i})^{Ag_i}) * p_{scale} \tag{3}$$

where Ag_i is the current number of Ag copies remaining, p_{scale} is an empirical scaling parameter, and l is defined as:

$$l = 1 - \frac{1}{1 + \frac{Ag_i}{Ag_0} * e^{-e_{scale}(E_{ij} - E_a)}} \tag{4}$$

with Ag_0 being the initial number of Ag copies, E_{ij} is the ΔG of clone j on cycle i, E_a is the activation energy, and e_{scale} is a pseudo-inverse temperature. The value of e_{scale} was set to 0.9 $\frac{kcal}{mol}$ to ensure BCRs were capable of capturing adequate antigen, while simultaneously ensuring a single immunization would produce primarily low-breadth antibodies–mimicking the finding that bnAbs often require many evolutionary cycles to evolve [16, 51].

For simplicity, we assume B cells have 100 BCRs on their surface with two Fab domains per receptor–resulting in 200 potential binding events per B cell (n_{max}). We assume each Fab has one opportunity to recognize and bind the Ag displayed by the FDC, defined by $P_{capture}$. Thus, we perform 200 independent Bernoulli trials for each BCR per GC cycle. The number of antigen copies captured by each B cell, in turn, determines how many divisions each population will undergo at the start of the following GC cycle when the B cells "return" to the dark zone via:

$$divisions = round(div_{max} \frac{R}{\frac{1}{A * n_{cap} + A} + R} + B) \tag{5}$$

Here, div_{max} is set to six to replicate experimental findings [21], R is the ratio of the number of antigens captured, n_{cap}, and n_{max}, and both A and B are empirical scaling parameters. The value obtained is then rounded to the nearest whole number.

After the antigen capture process, the B cells then move on to seeking T cell help. We assume the top 70% (w.r.t. ΔG) are capable of finding and soliciting help from the T_H cell population. From here, B cells may differentiate into memory cells (B_{mem} cells) or plasma cells (B_{plasma} cells) with a probability of 30%. The remaining 70% are then recycled to continue the maturation process. During each cycle, sequences describing the plasma B cells exiting the GC are then written to a log file.

5 Empirical Evaluation

To evaluate the efficacy of our proposed pipeline, we perform an experiment aiming at improving an HIV antigen. Our base antigen is the BG505 SOSIP.664 (BG505) gp120 sequence [41]. BG505 has been shown previously to both stably express in the native trimeric form [27] and elicit a productive immune response in humans and a variety of animal models [41]. Further, BG505 has been used previously in the design of anti-CD4bs antigens for potential vaccine regimens [11]. The antigens presented in Conti et al. [11], EU577271 (EU), HQ217523 (HQ), and KR423280 (KR), provide additional baseline antigens to compare against the output of the computational pipeline presented here. Our experiment evaluates our pipeline by letting it improve BG505, simulating a one-shot vaccination protocol using the designed antigen, then computing the binding strength of the resulting antibody population against an experimentally derived panel of diverse, circulating HIV viruses [13]. A successful HIV vaccine must elicit high-breadth antibodies, therefore we are seeking "generalist" antigens, which elicit improvement in the binding affinity across the entire panel, rather than strain- or subtype-specific responses to certain viruses. Our chosen input antibody is the germline of the broadly neutralizing VRC01, which in the literature has both been isolated in immune patients and had its germline antibody identified [53, 59].

To calculate the free energy of our complexes, we utilized the PRODIGY implementation in the high-throughput package, ppdx [10, 54]. To further increase the efficiency of our simulations, we elected to compute all binding affinities using the sum of single-points approach whereby the effects of a multi-point mutation is estimated as a linear combination of the single-point mutational impacts [26]. Estimates of the binding affinity were obtained via homology modeling all possible substitutions onto a known structure (PDB: 5FYJ [49]) individually. The resulting table of $\Delta\Delta G$ values for the GL-BG505 complex was then used to compute the scores of each antigen construct designed by the DSO agent. The optimization process was performed using a batch size of 1,000 sequences for 1,000 iterations, generating a total of 1,000,000 samples.

To narrow the mutational landscape being explored by the DSO agent, we began by defining the epitope-paratope interface using a simple distance cutoff of 10A between gp120 (antigen) and VRC01 (antibody)–reducing the number of mutable positions from approx. 400 to 76. To increase the likelihood the mutations introduced to the antigen do not disrupt the glycosylation on the protein surface, we limited the mutations around any potential N-linked glycosylation site (PNGS) to further reduce the number of mutable positions to 59. Finally, we limited the DSO agent to a maximum of three simultaneous mutations, and did not allow mutations to cysteine, proline, or tryptophan to

prevent the formation of new disulfide bonds and due to a bias toward large, bulky, and hydrophobic residues in our scoring function. After obtaining the DSO designed antigens, we selected ten diverse sequences (w.r.t. position and residue) for further evaluation in the affinity maturation simulation. In Table 1, we present the results from the top-performing design in the affinity maturation simulation from this initial set of ten.

The antigen-antibody interface is shown in Figure 2, with the residues within the distance cutoff highlighted in yellow. The residues mutated (HXB2 numbering: R191D, E379V, and Q437Y) in our antigen design, BV_1, are shown in the ball-and-stick representation. We hypothesize the improvement in binding affinity arises from improved charge-charge interactions by removing an acidic residue buried at the interface (E379V) and by inverting the charge on the periphery of the complementarity-determining regions (CDRs) of the antibody (R191D). Additionally, Q437Y appears to aid by providing a large, polar surface capable of stabilizing both hydrogen bond networks and preventing the water layer surrounding the protein from penetrating into the hydrophobic core. Further molecular dynamics simulations and *in vitro* experiments are warranted to characterize the impacts these mutations have on the structure and stability of the gp120 antigen, as well as the antigen-antibody complex.

Table 1 shows the improvement in potency of the antibody response over the germline upon vaccination with a given antigen (BG505, EU, HQ, KR, or BV_1). Values were calculated against the global panel [13], where each row in the table corresponds a given panel viral strain with their accession numbers listed. For the protocol using the parental BG505 sequence, the values presented represent the $\Delta\Delta G$ [$\frac{kcal}{mol}$] of the evolved antibody response, compared to the GL affinity. As the remaining antigens are variants of BG505, the values shown in Table 1 represent the percent change in $\Delta\Delta G$ relative to BG505.

There is an improvement in the binding strength to all of the panel sequences compared to BG505 when using BV_1, with a maximum improvement of 28%. Notably, our antigen outperforms the other computationally designed antigens in binding to each of the panel sequences but one (FJ444437), where both KR and BV_1 showed 25% and 21% improvement, respectively. In these experiments, BV_1 is clearly the top overall performing antigen with respect to binding improvement to diverse panel viruses. We expect this will correlate to a substantial improvement in the neutralization capacity of the immune response against HIV viruses *in vivo*. Further, the improved immune response was

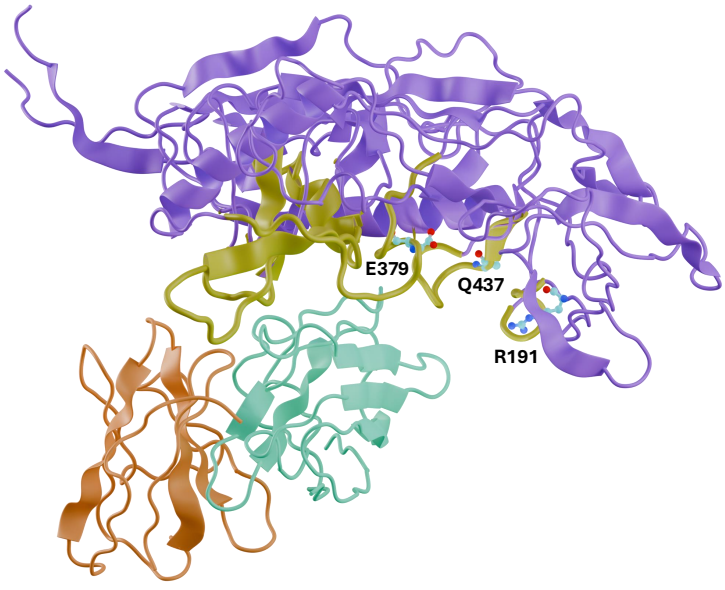

Figure 2: Crystal structure of gp120 (purple) in complex with VRC01 (V_H: blue; V_L: orange) with the region within the 10A cutoff highlighted (yellow) (PDB: 5FYJ [49]). The mutated positions in BV_1 (R191D, E379Y, Q437Y) are shown in the ball and stick representation.

Panel Ag	BG505	EU	HQ	KR	BV_1 (ours)
AY835445	0.753	-8%	4%	1%	16%
EF117261	0.502	-6%	3%	5%	24%
EF117271	0.997	9%	11%	14%	26%
FJ443575	1.116	1%	4%	1%	18%
FJ444437	0.787	11%	19%	25%	21%
FJ817366	0.739	-8%	-6%	4%	11%
FJ817370	0.688	-8%	-3%	2%	16%
HM215279	0.562	-11%	-7%	-2%	7%
HM215312	0.749	-10%	-3%	0%	12%
HM215364	0.559	7%	17%	14%	21%
HM215418	0.673	0%	10%	6%	28%
HM215427	0.798	-11%	-2%	1%	17%

Table 1: The average $\Delta\Delta G$ [$\frac{kcal}{mol}$] against a global HIV panel [13] inoculating with the BG505 parent sequence, and the percent change of the induced antibody response when using the BIOVAX designed antigen, BV_1, or previously optimized BG505 antigens (EU, HQ, KR) [cite], over the wildtype BG505 parental sequence. We show the best results per panel Ag in green and the worst results in red.

observed across almost all (11/12) viruses in the panel, indicating BV_1 may serve as a route to induce and promote the formation of bnAbs.

While further *in vivo* and *in vitro* validations of antigens developed by our pipeline is necessary to confirm the benefits of the specific antigens developed, this positive result showcases the power of our AI-based pipeline in assisting in the development of antigens for vaccination.

6 Conclusion

The development of better vaccine design pipelines is of utmost importance for public health, both because many pathogens (e.g., HIV) still elude effective vaccination, and because some existing vaccines either lack in breadth of variants covered or suffer from low protection effectiveness (e.g., influenza vaccines).

We here propose BIOVAX (**B**ioinformatics-**O**ptimized **V**accine **A**ntigen e**X**plorer), a computational AI-based pipeline to develop effective antigens for vaccination. Starting from a functional antigen, we perform Symbolic Optimization to design a diverse set of antigens that bind strongly to a broadly neutralizing antibody. Those designs are further evaluated in an affinity maturation simulation, and our final proposed design is the one with highest estimated breadth of protection. Further validation of the model outputs both *in vivo* and *in vitro* are warranted for the antibody repertoire and the binding characteristics therein, as well as considerations for other potential bnAb targets.

We perform an *in silico* evaluation of BIOVAX using real HIV targets and show that the designed antigen improves the binding of the estimated immune response for every single (real) virus in our test set when compared to the wildtype, and is a clear winner when compared against other computationally developed antigens from the literature. BIOVAX is a powerful tool to perform AI-assisted vaccine development.

Acknowledgments and Disclosure of Funding

The GUIDE program is executed by the Joint Program Executive Office for Chemical, Biological, Radiological and Nuclear Defense (JPEO-CBRND) Joint Project Lead for Enabling Biotechnologies (JPL CBRND EB) on behalf of the Department of Defense's Chemical and Biological Defense Program. The views expressed in this publication reflect the views of the authors and do not necessarily reflect the position of the Department of the Army, Department of Defense, nor the United States Government. References to non-federal entities do not constitute or imply Department of Defense or Army endorsement of any company or organization. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC. LLNL-CONF-869324.

References

[1] Christopher DC Allen et al. “Imaging of germinal center selection events during affinity maturation”. In: *Science* 315.5811 (2007), pp. 528–531.

[2] Frances H Arnold. “Design by directed evolution”. In: *Accounts of Chemical Research* 31.3 (1998), pp. 125–131.

[3] Lindsey R Baden et al. “Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine”. In: *New England Journal of Medicine* 384.5 (2021), pp. 403–416.

[4] Minkyung Baek et al. “Accurate prediction of protein structures and interactions using a three-track neural network”. In: *Science* 373.6557 (2021), pp. 871–876.

[5] Riyue Bao et al. “Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing”. In: *Cancer Informatics* 13 (2014), CIN–S13779.

[6] Kyle A Barlow et al. “Flex ddG: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation”. In: *The Journal of Physical Chemistry B* 122.21 (2018), pp. 5389–5399.

[7] Claudia Berek and Cesar Milstein. “Mutation drift and repertoire shift in the maturation of the immune response”. In: *Immunological Reviews* 96.1 (1987), pp. 23–41.

[8] Muhammad Bilal et al. “State-of-the-art protein engineering approaches using biological macromolecules: A review from immobilization to implementation view point”. In: *International Journal of Biological Macromolecules* 108 (2018), pp. 893–901.

[9] Liang Chen et al. “Trends in the development of miRNA bioinformatics tools”. In: *Briefings in Bioinformatics* 20.5 (2019), pp. 1836–1852.

[10] Simone Conti, Victor Ovchinnikov, and Martin Karplus. “ppdx: Automated modeling of protein–protein interaction descriptors for use with machine learning”. In: *Journal of Computational Chemistry* 43.25 (2022), pp. 1747–1757.

[11] Simone Conti et al. “Design of immunogens to elicit broadly neutralizing antibodies against HIV targeting the CD4 binding site”. In: *Proceedings of the National Academy of Sciences* 118.9 (2021), e2018338118.

[12] Simone Conti et al. “Multiscale affinity maturation simulations to elicit broadly neutralizing antibodies against HIV”. In: *PLoS Computational Biology* 18.4 (2022), e1009391.

[13] Allan deCamp et al. “Global panel of HIV-1 Env reference strains for standardized assessments of vaccine-elicited neutralizing antibodies”. In: *Journal of Virology* 88.5 (2014), pp. 2489–2507.

[14] Thomas A Desautels et al. “Computationally restoring the potency of a clinical antibody against Omicron”. In: *Nature* (2024), pp. 1–8.

[15] Jhon Ralph Enterina, Lanshi Wu, and Robert E Campbell. “Emerging fluorescent protein technologies”. In: *Current Opinion in Chemical Biology* 27 (2015), pp. 10–17.

[16] Amelia Escolano et al. “Sequential immunization elicits broadly neutralizing anti-HIV-1 antibodies in Ig knockin mice”. In: *Cell* 166.6 (2016), pp. 1445–1458.

[17] Jonathan G Faris et al. “Moving the needle: Employing deep reinforcement learning to push the boundaries of coarse-grained vaccine models”. In: *Frontiers in Immunology* 13 (2022), p. 1029167.

[18] Jonathan G Faris et al. “Pareto Front Training For Multi-Objective Symbolic Optimization”. In: *Adaptive and Learning Agents (ALA) Workshop at AAMAS*. 2024.

[19] Makda S Gebre et al. “Novel approaches for vaccine development”. In: *Cell* 184.6 (2021), pp. 1589–1603.

[20] Samik Ghosh et al. “Software for systems biology: from tools to integrated platforms”. In: *Nature Reviews Genetics* 12.12 (2011), pp. 821–832.

[21] Alexander D Gitlin, Ziv Shulman, and Michel C Nussenzweig. “Clonal selection in the germinal centre by regulated proliferation and hypermutation”. In: *Nature* 509.7502 (2014), pp. 637–640.

[22] Sarah A Griffith and Laura E McCoy. “To bnAb or not to bnAb: defining broadly neutralising antibodies against HIV-1”. In: *Frontiers in immunology* 12 (2021), p. 708227.

[23] Barton F Haynes et al. “Strategies for HIV-1 vaccines that induce broadly neutralizing antibodies”. In: *Nature Reviews Immunology* 23.3 (2023), pp. 142–158.

[24] Lun Hu et al. “A survey on computational models for predicting protein–protein interactions”. In: *Briefings in Bioinformatics* 22.5 (2021), bbab036.

[25] Katharina Imkeller and Hedda Wardemann. “Assessing human B cell repertoire diversity and convergence”. In: *Immunological Reviews* 284.1 (2018), pp. 51–66.

[26] Sherlyn Jemimah and M Michael Gromiha. “Exploring additivity effects of double mutations on the binding affinity of protein-protein complexes”. In: *Proteins: Structure, Function, and Bioinformatics* 86.5 (2018), pp. 536–547.

[27] Jean-Philippe Julien et al. “Crystal structure of a soluble cleaved HIV-1 envelope trimer”. In: *Science* 342.6165 (2013), pp. 1477–1483.

[28] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (2021), pp. 583–589.

[29] Kenneth Lundstrom. “Latest development on RNA-based drugs and vaccines”. In: *Future Science OA* 4.5 (2018), FSO300.

[30] Stefan Lutz. “Beyond directed evolution—semi-rational protein engineering and design”. In: *Current Opinion in Biotechnology* 21.6 (2010), pp. 734–743.

[31] Walter A Orenstein and Rafi Ahmed. *Simply put: Vaccination saves lives*. 2017.

[32] Victor Ovchinnikov et al. “Role of framework mutations and antibody flexibility in the evolution of broadly neutralizing antibodies”. In: *Elife* 7 (2018), e33038.

[33] Jesper Pallesen et al. “Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen”. In: *Proceedings of the National Academy of Sciences* 114.35 (2017), E7348–E7357.

[34] Bo Peng et al. “Replicating rather than nonreplicating adenovirus-human immunodeficiency virus recombinant vaccines are better at eliciting potent cellular immunity and priming high-titer antibodies”. In: *Journal of Virology* 79.16 (2005), pp. 10200–10209.

[35] Brenden K Petersen et al. “Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients”. In: *Proceeding of the International Conference on Learning Representations (ICLR)* (2021).

[36] Jacob F Pettit et al. “Learning sparse symbolic policies for sepsis treatment”. In: *Interpretable ML in Healthcare Workshop at ICML*. 2021.

[37] Fernando P Polack et al. “Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine”. In: *New England Journal of Medicine* 383.27 (2020), pp. 2603–2615.

[38] Renu Poria et al. “Vaccine development: Current trends and technologies”. In: *Life Sciences* (2023), p. 122331.

[39] Rino Rappuoli et al. “Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design”. In: *Journal of Experimental Medicine* 213.4 (2016), pp. 469–481.

[40] Colin A Russell et al. “Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses”. In: *Vaccine* 26 (2008), pp. D31–D34.

[41] Rogier W Sanders et al. “A next-generation cleaved, soluble HIV-1 Env trimer, BG505 SOSIP. 664 gp140, expresses multiple epitopes for broadly neutralizing but not non-neutralizing antibodies”. In: *PLoS Pathogens* 9.9 (2013), e1003618.

[42] Tanja A Schwickert et al. “In vivo imaging of germinal centres reveals a dynamic open structure”. In: *Nature* 446.7131 (2007), pp. 83–87.

[43] Mark J Shlomchik and Florian Weisel. “Germinal center selection and the development of memory B and plasma cells”. In: *Immunological Reviews* 247.1 (2012), pp. 52–63.

[44] Felipe Leno da Silva et al. “AutoTG: Reinforcement learning-based symbolic optimization for AI-assisted power converter design”. In: *IEEE Journal of Emerging and Selected Topics in Industrial Electronics* (2023).

[45] Felipe Leno da Silva et al. “Language model-accelerated deep symbolic optimization”. In: *Neural Computing and Applications* (2023), pp. 1–17.

[46] Felipe Leno da Silva et al. “Toward Multi-Fidelity Reinforcement Learning for Symbolic Optimization”. In: *Adaptive and Learning Agents (ALA) Workshop at AAMAS*. 2023.

[47] Ruth E Soria-Guerra et al. “An overview of bioinformatics tools for epitope prediction: implications on vaccine development”. In: *Journal of Biomedical Informatics* 53 (2015), pp. 405–414.

[48] Kayla G Sprenger et al. “Optimizing immunization protocols to elicit broadly neutralizing antibodies”. In: *Proceedings of the National Academy of Sciences* 117.33 (2020), pp. 20077–20087.

[49] Guillaume BE Stewart-Jones et al. “Trimeric HIV-1-Env structures define glycan shields from clades A, B, and G”. In: *Cell* 165.4 (2016), pp. 813–826.

[50] Gabriel D Victora and Michel C Nussenzweig. “Germinal centers”. In: *Annual Review of Immunology* 30.1 (2012), pp. 429–457.

[51] Shenshen Wang et al. “Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies”. In: *Cell* 160.4 (2015), pp. 785–797.

[52] Derek N Woolfson. “A brief history of de novo protein design: minimal, rational, and computational”. In: *Journal of Molecular Biology* 433.20 (2021), p. 167160.

[53] Xueling Wu et al. “Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1”. In: *Science* 329.5993 (2010), pp. 856–861.

[54] Li C Xue et al. “PRODIGY: a web server for predicting the binding affinity of protein–protein complexes”. In: *Bioinformatics* 32.23 (2016), pp. 3676–3678.

[55] Gur Yaari et al. “Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data”. In: *Frontiers in Immunology* 4 (2013), p. 358.

[56] Jian Ye et al. “IgBLAST: an immunoglobulin variable domain sequence analysis tool”. In: *Nucleic Acids Research* 41.W1 (2013), W34–W40.

[57] Jingshan Zhang and Eugene I Shakhnovich. “Optimality of mutation and selection in germinal centers”. In: *PLoS Computational Biology* 6.6 (2010), e1000800.

[58] Yifei Zhang, Jun Ge, and Zheng Liu. “Enhanced activity of immobilized or chemically modified enzymes”. In: *American Chemical Society Catalysis* 5.8 (2015), pp. 4503–4513.

[59] Tongqing Zhou et al. “Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies”. In: *Immunity* 39.2 (2013), pp. 245–258.