Game Theoretic Antibody Design

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Vaccines have had a marked impact on public health. However, designing vaccines is a laborious processes, involving extensive and expensive experimentation. Computational vaccine design promises to be game changing and commonly involves two steps: computational design of antibodies, followed by computational design of a vaccine which promotes generation of such antibodies. We focus on the first step: antibody design. An important challenge is the possibility of rapid viral mutations which escape antibody binding, as is the case with HIV and influenza. Indeed, in both these cases, evolution of the viral antigen has thus far foiled attempts at designing an effective long-term vaccine. A common way to capture viral evolution is to use a *fixed* panel of known viral variants, with the goal of designing a *broadly binding* antibody (i.e., one which binds most, or all of these). We propose a novel game theoretic approach to this problem, which allows us to capture not merely a fixed panel of viral variants, but also a combinatorial space of mutations from these. Our approach combines learning a linear approximation of binding stability energy of the antibody-virus complex with bi-level integer linear programming, which we transform into a single-level mixed-integer linear program. Through a series of simulation experiments we demonstrate the efficacy of our proposed approach.

1 Introduction

Infectious diseases pose a major threat to public health. In 2016, about 36.7 million people were living with HIV, and it resulted in 1 million deaths [14]. From the time AIDS was identified, it has caused an estimated 35 million deaths worldwide [19]. A recent Ebola outbreak in Africa killed thousands [3], and annual influenza outbreaks affect millions, with hundreds of thousands hospitalized, and thousands dying from the influenza or its side-effects [4]. Vaccination therapies are among the most important methods for combating infectious diseases. Vaccines are external substances that stimulate the immune system to produce antibodies that bind to the vaccine substance. Traditional vaccine design involves laborious and costly lab work aimed at finding just the right substance which would successfully and reliably elicit antibodies binding the target pathogen. Recently, a promising approach has been taking shape in which vaccines are designed *computationally*, making use of modern computational protein modeling tools, such as ROSETTA [1]. One of the common approaches involves two steps: first, finding an antibody with desired neutralization characteristics, and second, finding a vaccine which binds tightly to the desired antibody, thereby eliciting the associated target immune response. We focus on the first step of computational antibody design.

The central goal in computational antibody design is to find an antibody protein sequence which neutralizes the target pathogen by *binding* to it. When two proteins (such as an antibody and viral proteins) bind, they form a *complex*, which is a configuration minimizing the total energy of the two molecules. Binding is typically highly specific: a small change in the sequence can destabilize binding. However, binding a single fixed virus is often insufficient: for example, viruses such as HIV and flu have many strains, and an antibody which neutralizes one will often fail to neutralize another. An area of active research in antibody design, therefore, is to develop and characterize *broadly binding antibodies*, i.e., antibodies which effectively bind to many variants of the pathogen [10]. Nevertheless, as a pathogen evolves, it may well still escape neutralization; for example, HIV has an extremely high mutation rate [8].

We propose a radically different approach for computational antibody design in the context of rapidly mutating viruses: using a game theoretic (Stackelberg game) model for the interaction between the antibody and the virus. In this game, the antibody designer chooses an antibody sequence, while the virus aims to maximally destabilize binding to the resulting antibody, subject to a constraint on the number of mutations (this constraint captures the fact that such a mutation has to be sufficiently likely). This game can be formulated as a bi-level optimization problem; unfortunately, such a formulation is quite intractable. We address tractability in three steps: first, we learn a linear approximation of the antibody-virus binding score as a function of its sequence (including all pairwise interactions at the binding site); second, we formulate the optimal virus escape problem as an integer linear program; and third, after relaxing the integrality constraint in the virus escape program and taking its dual, we formulate the antibody design bi-level problem as a mixed-integer linear program. Our experimental results demonstrate that our approach is extremely effective against two recent prior approaches for HIV antibody design.

Related Work Conceptually, our work follows the research on game theoretic antibody design in [15], which extends the insights of Stackelberg security games [11] to the vaccine design domain. A major challenge is the enormous search space ($\sim 20^{60}$, with 30 binding sites on each protein and 20 possible amino acids) which is tackled using local search. Our contribution is a compact formulation to compute the optimal global solution. Since antibodies are protein sequences, our work relates significantly to computational protein design. Recent advances involve multi-specificity design to achieve protein design with respect to more than one targets [17]. The most relevant prior work is Breadth Optimization in Antibody Design (BROAD) that incorporates machine learning and sequence optimization for efficient sampling in the sequence space [18]. However, while BROAD maximizes breadth over an existing virus panel, our approach of game-theoretic antibody design goes significantly further as the designed antibody continues to bind against virus escape mutations. There have been numerous efforts in learning protein structure, function and interactions from sequence data, of which Kamisetty et al. [12] is the most relevant to our effort. More remotely related work include game theoretic models of vaccination decisions [2, 5, 13]. However, these model human decisions about being vaccinated, whereas our model involves molecular-level interactions between immunity and pathogen.

2 Game Theoretic Model and Solution Approach

We define an antibody or virus primary sequence as a sequence (vector) of amino acids as in previous work [15]. Let $\mathbf{c}$ denote the native virus (the initial virus strain before mutations) and $(\mathbf{a}, \mathbf{v})$ be arbitrary antibody and virus sequences respectively. Let $\mathcal{B}(\mathbf{a}, \mathbf{v})$ and $\mathcal{S}(\mathbf{a}, \mathbf{v})$ denote the binding energy and the thermodynamic stability scores of the antibody-virus complex. A combination of these is used as the overall *energy score* (often known as the z-score) of the complex, which is what we actually work with, and denote by $\mathcal{Z}(\mathbf{a}, \mathbf{v})$. Also, lower (more negative) scores indicate stronger binding and stability of the antibody-virus complex.

The virus sequence attempts to escape binding to the antibody by making a series of mutations. We can represent the number of mutations in a mutated virus sequence $\mathbf{v}$ from the native $\mathbf{c}$ as $\|\mathbf{v} - \mathbf{c}\|_0$, where the l_0 norm computes the number of sequence positions in $\mathbf{v}$ that are different from $\mathbf{c}$. Given an antibody $\mathbf{a}$, we model the virus as making up to α mutations with the goal of maximizing its binding energy score so as to destabilize binding. In general, there are many potential virus variants that can infect an individual. To capture this, we consider T virus sequences of different types t in a virus panel, each starting from a native sequence $\mathbf{c}^t$ and making mutations to escape binding to $\mathbf{a}$. We can formally represent the combined *best response* of the virus panel to a fixed antibody $\mathbf{a}$ as: $\underset{\mathbf{v}^t \in \mathcal{V}}{\text{maximize}} \sum_{t=1}^{T} \mathcal{Z}(\mathbf{a}, \mathbf{v}^t)$ subject to $\|\mathbf{v}^t - \mathbf{c}^t\|_0 = \alpha, \forall t$. where $\mathcal{V}$ is the space of virus sequences under consideration. The antibody designer's decision problem is to choose an antibody which minimizes the energy scores (strengthens binding and stability) with respect to the virus panel $\{1, \dots, T\}$, accounting for potential mutations of each virus in response. This gives rise to the following bi-level optimization problem for antibody design: $\min_{\mathbf{a} \in \mathcal{A}} \max_{\mathbf{v}^t \in \mathcal{V}} \sum_{t=1}^{T} \mathcal{Z}(\mathbf{a}, \mathbf{v}^t)$

subject to $\|\mathbf{v}^t - \mathbf{c}^t\|_0 = \alpha, \forall t$ where $\mathcal{A}$ is the antibody design space. Observe that the antibody-virus interaction in our model can be viewed as a Stackelberg game in which the designer (antibody) is the leader, and each virus is the follower, who chooses an alternative virus sequence in response to the antibody chosen by the designer. Moreover, this game is zero-sum: the designer minimizes the energy score, a quantity which is maximized by each virus t.

The bi-level optimization problem is intractable in general, even when simulated using the ROSETTA software. In particular, computing such a function using ROSETTA even for a given pair of sequences requires many runs of stochastic local search, and takes on the order of minutes or hours. We make progress by approximating the complex black-box ROSETTA energy function $\mathcal{Z}(\mathbf{a}, \mathbf{v})$ by a bi-linear function of the antibody and virus sequences, similar to the approach proposed by Kamisetty et al. [12]. The model is based on an assumption that the binding and stability of an antibody-virus complex is primarily determined by two factors: a) the individual amino acids in each binding position of the antibody and the virus respectively, and b) the effects of the pairwise amino acid interactions between the antibody and the virus. Using this model, we formulate the virus optimal escape problem as an integer linear program (ILP). However, our underlying problem of antibody design is still a bi-level problem with integer variables, which in general, is extremely challenging to solve. At the high level, we propose to leverage the linear structure of the problem to solve it. First, we relax the integrality constraint of the inner (virus escape) problem and show that the resulting relaxed LP has integral optimal solutions. Next, we obtain the dual of the relaxed LP which we embed into the outer integer linear program. By relaxation, combined with strong duality of linear programming, the resulting mixed-integer linear program (MILP) minimizes an *upper bound* on the z-score objective with respect to optimal virus escape.

3 Experiments

The data comprises the anti-HIV antibody VRC23 [9] (the native antibody) against a set of 180 diverse HIV gp120 virus sequences (derived from Chuang et al. [6]). To generate our training data, we make random antibody and virus substitutions in the binding sites of VRC23 and the set of 180 virus sequences ($N_a = 27$ and $N_v = 32$). Each antibody/virus variant has five randomly selected amino acid mutations. All antibody-virus pairs are subjected to an energy minimization via the ROSETTA relax protocol [7]). We generate 50 models of each antibody-virus pair and choose the lowest scoring model in each case. We then construct the dataset for our experiments with a total of 7360 such random antibody-virus combinations (including VRC23 and the 180 virus sequences).

First, we learn the bi-linear z-score model, using sparse matrices to represent the feature space and use the Lasso implementation in scikit-learn [16] with l_1 (sparse) regularization. To measure the accuracy of predictions, we compute the correlation coefficient between the ROSETTA computed z-scores and the scores predicted by regression. We perform a 10-fold cross validation experiment with 80% of the data for training and 20% for testing. Based on this parameter tuning, we choose regularization parameter $\lambda = 0.01$ with an average correlation of 0.85 between the predicted and the ROSETTA computed z-scores. We denote our proposed antibody design approach as STRONG: STackelberg game theoretic model for RObust aNtibody desiGn.

Comparison against previous work BROAD [18] is a state of the art algorithm for antibody design against a *fixed* panel of HIV virus variants that involves generating a large training set of binding and stability scores using ROSETTA, fitting linear models to predict binding and stability, and solving an ILP to compute an optimal broadly binding antibody sequence.

We perform the comparison following the experimental workflow in BROAD. We construct 50 random subsamples of the full training data corresponding to $T = 100$ out of the 180 virus sequences We train binding and stability prediction models on this data and compute the BROAD antibody sequence by solving an ILP with the T virus sequences in the training subsample. Next, for each training subsample we learn the bi-linear model and save the coefficients. Then, we solve the antibody design MILP to compute the corresponding STRONG antibody for a given α. Given this antibody, we solve the virus escape ILP to compute T escaping virus sequences corresponding to each of the T training sequences (native). We use CPLEX version 12.51 to solve the (mixed) integer linear programs. Finally, we train a z-score model on the full dataset ($T = 180$). We evaluate the BROAD and the STRONG antibody sequences in terms of the predicted z-score against a) the full 180 virus panel and b) the 100 escaping virus sequences in case of each training subsample. This procedure is outlined in Algorithm 1. As we show in Figure 1 STRONG is significantly better in minimizing the z-score objective as compared to BROAD.

Algorithm 1 Generating and evaluating STRONG antibody candidates

```
for each random training set subsample corresponding to $T = 100$ virus sequences do
    Training Data: $\mathcal{B}(\mathbf{a}, \mathbf{v}), \mathcal{S}(\mathbf{a}, \mathbf{v}), \mathcal{Z}(\mathbf{a}, \mathbf{v})$ corresponding to the $T$ training sequences
    Learning: bi-linear model for z-score
    Optimization: STRONG antibody $\leftarrow$ bi-level MILP, escaping set $\leftarrow$ virus escape ILP
    Evaluation: predicted z-score using model trained on the full dataset, and ROSETTA modeling
```

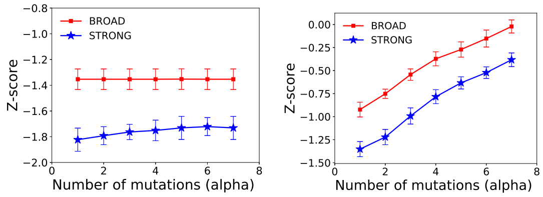

Figure 1: Comparison between STRONG and BROAD in terms of the z-score objective (lower is better): on the full 180 virus panel (left) and the 180 escaping virus set (right).

Finally, we evaluate in terms of the breadth of binding (fraction of viruses in the evaluation panel to which the designed antibody binds) generated using ROSETTA structure modeling. We choose 50 random subsamples of training sets with $T = 30$ virus sequences. Based on binding and stability models trained on the full dataset, we generate the top 10 BROAD candidates. Next, we generate the STRONG antibody for a randomly chosen top BROAD candidate using $\alpha = 5$. We perform ROSETTA structure modeling on these antibody candidates (one BROAD and one STRONG candidate) and the escaping set of 30 virus sequences. For comparison, we also include the native antibody VRC23. We present the ROSETTA computed breadth in each case in Table 1. STRONG significantly outperforms BROAD against the escaping virus panel while it continues to be effective against the training panel.

We also compare with the game-theoretic antibody design approach in [15]. The proposed approach is significantly better in minimizing the objective (z-score).

Virus Sequences for Evaluation	VRC23	BROAD	STRONG
180 HIV panel	53.3	100	96.1
30 Escaping virus sequences	43.3	86.7	93.3
30 Training virus sequences	56.7	100	100

Table 1: ROSETTA structure modeling results: breadth of binding (%).

4 Discussions

We proposed an efficient approach for computational antibody design using a Stackelberg game model for the interaction between the antibody and the virus. We formulated the game as a bi-level optimization problem, and proposed a method for solving it by leveraging a bi-linear model predicting binding stability as a function of antibody and virus sequence, combined with integer programming. Our experiments show that our approach significantly outperforms both the prior game theoretic alternative, and a state-of-the-art broadly binding antibody design algorithm.

While the proposed framework was developed in the context of antibody design (vaccination) for HIV viruses, it is applicable directly to vaccination design for other viruses (e.g., influenza). It can also be applied almost without change to drug design with the search over relevant drug fragment space. Additionally, this framework is not unique to viruses, and can be generalized directly to other pathogens (e.g., ebola vaccine design) as well as antibiotic treatments for many bacterial infections (e.g., to address antibiotic resistance).

References

[1] Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.

[2] Chris T. Bauch and David J.D. Earn. Vaccination and the theory of games. *Proceedings of the National Academy of Sciences*, 101(36):13391–13394, 2004.

[3] CDC. 2014 ebola outbreak in west africa, 2014. http://www.cdc.gov/vhf/ebola/outbreaks/guinea/.

[4] CDC. Seasonal influenza, more information, 2018. https://www.cdc.gov/flu/about/qa/disease.htm.

[5] GB Chapman, M Li, J Vietri, Y Ibuka, D Thomas, H Yoon, and AP. Galvani. Using game theory to examine incentives in influenza vaccination behavior. *Psychological Science*, 23(9):1008–1015, 2012.

[6] Gwo-Yu Chuang, Priyamvada Acharya, Stephen D Schmidt, Yongping Yang, Mark K Louder, Tongqing Zhou, Young Do Kwon, Marie Pancera, Robert T Bailer, Nicole A Doria-Rose, et al. Residue-level prediction of hiv-1 antibody epitopes based on neutralization of diverse viral strains. *Journal of virology*, 87(18):10047–10058, 2013.

[7] Steven A Combs, Samuel L DeLuca, Stephanie H DeLuca, Gordon H Lemmon, David P Nannemann, Elizabeth D Nguyen, Jordan R Willis, Jonathan H Sheehan, and Jens Meiler. Small-molecule ligand docking into comparative models with rosetta. *Nature protocols*, 8(7):1277, 2013.

[8] José M Cuevas, Ron Geller, Raquel Garijo, José López-Aldeguer, and Rafael Sanjuán. Extremely high mutation rate of hiv-1 in vivo. *PLoS biology*, 13(9):e1002251, 2015.

[9] Ivelin S Georgiev, Nicole A Doria-Rose, Tongqing Zhou, Young Do Kwon, Ryan P Staupe, Stephanie Moquin, Gwo-Yu Chuang, Mark K Louder, Stephen D Schmidt, Han R Altae-Tran, et al. Delineating antibody recognition in polyclonal sera from patterns of hiv-1 isolate neutralization. *Science*, 340(6133):751–756, 2013.

[10] Jinghe Huang, Byong H. Kang, Marie Pancera, Jeong Hyun Lee, Tommy Tong, Yu Feng, Ivelin S. Georgiev, Gwo-Yu Chuang, Aliaksandr Druz, Nicole A. Doria-Rose, Leo Laub, Kwinten Sliepen, Marit J. van Gils, Alba Torrents de la Pena, Ronald Derking, Per-Johan Klasse, Stephen A. Migueles, Robert T. Bailer, Munir Alam, Pavel Pugach, Barton F. Haynes, Richard T. Wyatt, Rogier W. Sanders, James M. Binley, and Andrew B. Ward. Broad and potent hiv-1 neutralization by a human antibody that binds the gp41-gp120 interface. *Nature*, 2014.

[11] Manish Jain, James Pita, Milind Tambe, Fernando Ordóñez, Praveen Paruchuri, and Sarit Kraus. Bayesian stackelberg games and their application for security at los angeles international airport. *SIGecom Exch.*, 7:10:1–10:3, June 2008.

[12] Hetunandan Kamisetty, Bornika Ghosh, Christopher James Langmead, and Chris Bailey-Kellogg. Learning sequence determinants of protein: Protein interaction specificity with sparse graphical models. *Journal of Computational Biology*, 22(6):474–486, 2015.

[13] Jingzhou Liu, Beth F. Kochin, Yonas I. Tekle, and Alison P. Galvani. Epidemiological game-theory dynamics of chickenpox vaccination in the usa and israel. *Journal of the Royal Society Interface*, 9(66):68–76, 2012.

[14] Joint United Nations Programme on HIV/AIDS et al. Fact sheet—latest statistics on the status of the aids epidemic, 2017.

[15] Swetasudha Panda and Yevgeniy Vorobeychik. Stackelberg games for vaccine design. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 1391–1399, 2015.

[16] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[17] Alexander M Sevy, Tim M Jacobs, James E Crowe Jr, and Jens Meiler. Design of protein multi-specificity using an independent sequence search reduces the barrier to low energy sequences. *PLoS computational biology*, 11(7):e1004300, 2015.

[18] Alexander M Sevy, Swetasudha Panda, James E Crowe Jr, Jens Meiler, and Yevgeniy Vorobeychik. Integrating linear optimization with structural modeling to increase hiv neutralization breadth. *PLoS computational biology*, 14(2):e1005999, 2018.

[19] UNAIDS. Fact sheet-latest statistics on the status of the aids epidemic, 2016.

5 Appendix

5.1 A Bi-Linear Representation of Energy Scores

We represent an antibody sequence $\mathbf{a}$ as a binary position by amino-acid matrix, with $a_{ij} = 1$ iff amino acid j appears in position i, and $a_{ij} = 0$ otherwise. Thus, $\sum_j a_{ij} = 1$, since exactly 1 amino acid can be in a given position. Similarly, virus protein sequence is represented as a binary matrix v_{ij} which is 1 iff amino acid j is in position i. Let N_a and N_v denote the number of binding positions on the antibody and the virus respectively, and let $M = 20$ denote the number of amino acids.

Amino acid contributions to the energy score can be modeled as a bipartite graph in which nodes represent the amino acids and the edges represent the pairwise amino acid interactions. Each antibody position node i has an associated weight vector $\mathbf{x}_i \in \mathbb{R}^M$. Similarly, each virus position node j has an associated weight vector $\mathbf{y}_j \in \mathbb{R}^M$. The edge (i, j) between antibody position node i and virus position node j has an associated weight matrix $Q_{ij} \in \mathbb{R}^{M\times M}$ to represent the position specific contribution to the energy score for each amino acid pair. Consequently, given $\mathbf{a}$ and $\mathbf{v}$, the energy score varies as the sum of individual amino acids and pairwise interaction effects. Given this setting, the z-score for a given pair $\mathbf{a}$ and $\mathbf{v}$ is defined as:

$$\mathcal{Z}(\mathbf{a}, \mathbf{v}) = \sum_{i=1}^{N_a}\sum_{j=1}^{M} x_{ij}a_{ij} + \sum_{i=1}^{N_v}\sum_{j=1}^{M} y_{ij}v_{ij} + \sum_{k=1}^{N_a}\sum_{l=1}^{N_v}\sum_{u=1}^{M}\sum_{m=1}^{M} a_{ku}q_{kl}^{um}v_{lm} + I \tag{1}$$

where I is the intercept term and q_{kl}^{um} represents $Q_{kl}(u, m)$.

Our bi-linear model thus has four sets of parameters: $\mathbf{x}_i$, $\mathbf{y}_j$, and Q_{ij} for all pairs of antibody and virus positions, i and j, respectively, and the intercept I. We learn these parameters by generating a dataset of ROSETTA energy function values for a number of pairs of antibody and virus sequences (as detailed in the experiments).

Armed with the bi-linear model described in this section, we can convert the hard bilevel optimization problem into a significantly more tractable mixed-integer linear program through a combination of convex relaxation and duality, as we describe next.

5.2 Integer Linear Program for Virus Escape

Our first step is to formulate the virus optimal escape problem as an integer linear program.

We start by observing that the number of mutations α can be computed using a dot product with the sequence representation described above. Specifically, $\mathbf{v}^t \cdot \mathbf{v}^t = N_v$ and $\mathbf{v}^t \cdot \mathbf{c}^t = N_v - \alpha$. Moreover, $\mathcal{Z}(\mathbf{a}, \mathbf{v})$ is now a linear function with the above sequence representation. These observations allow us to formulate the virus escape optimization as an integer linear program (ILP). Since in this problem the antibody $\mathbf{a}$ is fixed, we can group the model in Equation 1 in terms of the variables $\mathbf{v}$ as $\sum_{i=1}^{N}\sum_{j=1}^{M} x_{ij}a_{ij} + \sum_{i=1}^{N}\sum_{j=1}^{M}\Big(y_{ij} + \sum_{k=1}^{N}\sum_{u=1}^{M} a_{ku}q_{uj}^{ki}\Big)v_{ij} + I$. Thus, the virus escape ILP for a particular native virus indexed by t (from a collection of T of these) can be formulated as follows:

$$\begin{aligned}
\underset{\mathbf{v}^t\in\mathcal{V}}{\text{maximize}}\quad & \sum_{t=1}^{T}\sum_{i=1}^{N_v}\sum_{j=1}^{M}\Big(y_{ij} + \sum_{k=1}^{N_a}\sum_{u=1}^{M} a_{ku}q_{uj}^{ki}\Big)v_{ij}^t + T\sum_{i=1}^{N_a}\sum_{j=1}^{M} x_{ij}a_{ij} \\
\text{subject to}\quad & \sum_{j=1}^{M} v_{ij}^t = 1, \forall i, t && \text{(2a)} \\
& N_v - \sum_{i=1}^{N_v}\sum_{j=1}^{M} v_{ij}^t c_{ij}^t = \alpha, \forall t && \text{(2b)} \\
& v_{ij}^t \leq L(p_{ij} - \theta), \forall i, j, t && \text{(2c)} \\
& v_{ij}^t \in \{0, 1\}, \forall i, j, t
\end{aligned}$$

where constraint 2a enforces that the binary variables v_{ij}^t at each antibody binding position should sum to 1, i.e., each position admits one amino acid. The term $\sum_{i=1}^{N_v}\sum_{j=1}^{M} v_{ij}^t c_{ij}^t$ in constraint 2b computes the dot product $\mathbf{v}^t \cdot \mathbf{c}^t$. The constraint 2c encodes the constraint that we only allow mutations at positions to amino acids which have been observed at a frequency $p_{ij} \geq \theta$ as a linear constraint; here, L is a large number.

5.3 Mixed Integer Linear Program for Antibody Design

While we can represent the optimization problem faced by the virus *given a fixed antibody* using a linear integer program, our underlying problem of antibody design is still a bi-level problem. Such bi-level problems (with integer variables, as in our case) are, in general, extremely challenging to solve.

At the high level, we propose to leverage the linear structure of the problem to solve it. First, we relax the integrality constraint of the inner (virus escape) problem. This yields a linear program, the dual of which we embed into the outer integer linear program. By relaxation, combined with strong duality of linear programming, the resulting mixed-integer linear program minimizes an *upper bound* on the z-score objective with respect to optimal virus escape.

We start with the ILP 2 computing the optimal virus escape, and relax the integrality constraint; that is, we relax the binary v_{ij}^t variables to be continuous and add the constraints $0 \leq v_{ij}^t \leq 1$. Next, we show that the resulting relaxed LP has integral optimal solutions.

We observe that the primal relaxed LP is feasible and bounded, and, therefore, the dual is also feasible and bounded, and (by strong duality) has the same solution as the primal. Let the associated (non-negative) dual variables be denoted by ψ_{ij}^t for each of the constraints $v_{ij}^t \leq 1$, and let ϕ_i^t (unrestricted), π^t (unrestricted) and ρ_{ij}^t (non-negative) denote the dual variables corresponding to constraints 2a, 2b, and 2c. Note that all dual variables are continuous. The dual LP is the given by the following ($\mathbf{a}$ is fixed here as in the primal LP):

$$
\begin{aligned}
\underset{\phi,\psi,\rho,\pi}{\text{minimize}} \; & \sum_{t=1}^{T}\left[\sum_{i=1}^{N_v}\phi_i^t-(N_v-\alpha)\pi^t+\sum_{i=1}^{N_a}\sum_{j=1}^{M}L(p_{ij}-\theta)\rho_{ij}^t+\sum_{i=1}^{N_v}\sum_{j=1}^{M}\psi_{ij}^t\right]+T\sum_{i=1}^{N_a}\sum_{j=1}^{M}x_{ij}a_{ij} \\
\text{subject to } & \phi_i^t-\pi^t c_{ij}^t+\rho_{ij}^t+\psi_{ij}^t \geq \left(y_{ij}+\sum_{k=1}^{N}\sum_{u=1}^{M}a_{ku}q_{uj}^{ki}\right), \forall i,j,t \qquad \text{(3a)} \\
& \psi,\rho \geq 0, \pi,\phi \text{ unrestricted variables}
\end{aligned}
$$

Next, we integrate this dual LP into the antibody optimization problem to formulate the following mixed integer linear program (MILP):

$$
\begin{aligned}
\underset{\mathbf{a}\in\mathcal{A},\phi,\psi,\rho,\pi}{\text{minimize}} \; & \sum_{t=1}^{T}\left[\sum_{i=1}^{N_v}\phi_i^t-(N_v-\alpha)\pi^t+\sum_{i=1}^{N_a}\sum_{j=1}^{M}L(p_{ij}-\theta)\rho_{ij}^t+\sum_{i=1}^{N_v}\sum_{j=1}^{M}\psi_{ij}^t\right]+T\sum_{i=1}^{N_a}\sum_{j=1}^{M}x_{ij}a_{ij} \\
\text{subject to } & \phi_i^t-\pi^t c_{ij}^t+\rho_{ij}^t+\psi_{ij}^t-\sum_{k=1}^{N}\sum_{u=1}^{M}a_{ku}q_{uj}^{ki} \geq y_{ij}, \forall i,j,t \qquad \text{(4a)} \\
& \sum_{u=1}^{M}a_{ku}=1, \forall u \qquad \text{(4b)} \\
& a_{ij}^t \in \{0,1\}, \forall i,j,t \\
& \psi,\rho \geq 0, \pi,\phi \text{ unrestricted variables}
\end{aligned}
$$

The variables now include the binary antibody variables a_{ij}, and the constraints ensure that these sum to 1 at each antibody binding position, i.e., each position admits one amino acid. An important observation we can make is that while originally we had bi-linear terms involving antibody and virus decision variables, these are decoupled after taking the dual, resulting in solely linear terms.