ExaLearn COVID-19 Pipeline Optimization

Byung-Jun Yoon[1,2]

[1]Computational Science Initiative, Brookhaven National Laboratory
[2]Department of Electrical and Computer Engineering, Texas A&M University

1 The Screening Pipeline

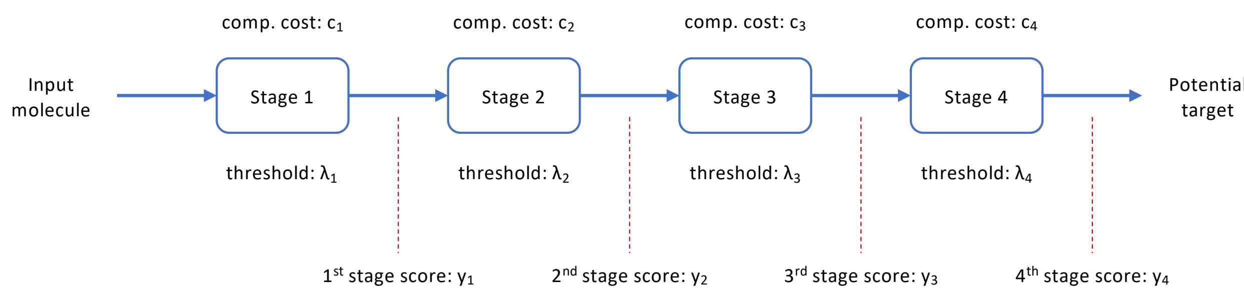

Figure 1: COVID-19 pipeline.

The pipeline for identifying potential targets is shown in Figure 1. For an input molecule $x \in \mathbb{X}$, we denote the score from the stage-i filter as $y_i = f_i(x)$. At stage-i, the molecule x is passed to the next stage if $y_i \geq \lambda_i$ for some threshold λ_i. Otherwise, it is discarded and not considered any further. We denote the set of molecules that pass the stage-i filter as $\mathbb{X}_i$,

$$\mathbb{X}_i = \{x \mid x \in \mathbb{X}_{i-1} \text{ and } f_i(x) \geq \lambda_i\} \tag{1.1}$$

where we denote $\mathbb{X}_0 \triangleq \mathbb{X}$ for convenience. Let us consider the joint distribution of the stage-1 to stage-4 scores, where the joint PDF is denoted by $f(y_1, y_2, y_3, y_4)$.

2 The Optimal Screening Strategy

The overall goal is to maximize the proportion (or number) of the potential target molecules that pass the stage-4 filter, which are promising target molecules whose stage-4 score exceeds the specified threshold ($f_4(x) \geq \lambda_4$). Suppose we choose a thresholding strategy $\psi = (\lambda_1, \lambda_2, \lambda_3)$, where we assume λ_4 is fixed based on criteria used to select the final set of potential targets. This will result in the following reward:

$$r(\psi) = r(\lambda_1, \lambda_2, \lambda_3) = \int_{\lambda_4}^{\infty} \int_{\lambda_3}^{\infty} \int_{\lambda_2}^{\infty} \int_{\lambda_1}^{\infty} f(y_1, y_2, y_3, y_4) dy_1 dy_2 dy_3 dy_4 \tag{2.1}$$

Equivalently, we can define the cost as follows:

$$c(\psi) = c(\lambda_1, \lambda_2, \lambda_3) = 1 - r(\lambda_1, \lambda_2, \lambda_3). \tag{2.2}$$

This leads to the constrained optimzation problem below:

$$\begin{aligned} \min_{\psi} \quad & c(\psi) \\ \text{s.t.} \quad & c_1|\mathbb{X}_0| + c_2|\mathbb{X}_1| + c_3|\mathbb{X}_2| + c_4|\mathbb{X}_3| \leq C \end{aligned} \tag{2.3}$$

where C is the total computational budget. We could also rewrite the constraint on the computational budget as follows:

$$c_2 \int_{\lambda_1}^{\infty} f(y_1) dy_1 + c_3 \int_{\lambda_2}^{\infty} \int_{\lambda_1}^{\infty} f(y_1, y_2) dy_1 dy_2 + c_4 \int_{\lambda_3}^{\infty} \int_{\lambda_2}^{\infty} \int_{\lambda_1}^{\infty} f(y_1, y_2, y_3) dy_1 dy_2 dy_3 \leq \frac{1}{|\mathbb{X}_0|} C - c_1 \tag{2.4}$$

based on the joint PDF $f(y_1, y_2, y_3, y_4)$.

3 MOCU and OED

However, we may not have no (or little) knowledge about the joint score distribution $f(y_1, y_2, y_3, y_4)$, which makes it impossible to choose the optimal strategy ψ that minimizes the cost given the constraints. Suppose the joint PDF $f_\theta(\cdot)$ can be parameterized by $\theta \in \Theta$, where we have its prior distribution $\pi(\theta)$. For a given θ, we define $\psi_\theta = \arg\min_\psi c_\theta(\psi)$ to be the optimal strategy that minimizes the cost given the constraints, where the cost $c_\theta(\psi)$ also depends on the joint score distribution $f_\theta(y_1, y_2, y_3, y_4)$. Now, we can compute the MOCU as follows:

$$M(\Theta) = E_\theta \Big[c_\theta(\psi^*) - c_\theta(\psi_\theta) \Big] \tag{3.1}$$

where $\psi^* = \arg\min_\psi E_\theta[c_\theta(\psi)]$.

Potential research directions In order to optimize the screening pipeline, we can use the above expression of MOCU to predict and compare the efficacy of potential "experiments" for reducing the uncertainty class Θ of the joint score distribution $f_\theta(\cdot)$.