# 1 Introduction

# 2 Problem Statement and Application

We first describe the application that motivates our research, and then we provide mathematical formalism to address a more general problem. In the last sub-section we derive our method in solving this problem.

## 2.1 Motivating application

We have two enzymes (Sfp from *Bacillus subtilis*, and PaAcpH from *Pseudomonas aeruginosa*), and a collection of peptides that can potentially act as a substrate for one or both of these enzymes. Our goal is to find a peptide that acts as a substrate for both of these enzymes, and is as short as possible.

To support this goal, we can do lab experiments, in which we synthesize a peptide and test, for each enzyme, whether it is a substrate. We need to find a policy that suggests which peptide to synthesize and test next, so as to reach our goal with as few experiments as possible.

Experiments have parallel setup, thus can be done with a batch of peptides at a time, and so the algorithm suggests a batch of peptides at a time, waiting for the results from the experiment before suggesting the next batch of peptides. A large collection of peptides would be considered by the algorithm for potential synthesis and testing, e.g., all peptides with length less than a given threshold. That is, we would consider more peptides than just those that are sub-peptides of peptides from the literature known to be substrates for one enzyme.

## 2.2 General Problem Statement

We now formalize and generalize our problem as an active learning problem, which includes but is not limited to our motivating application.

Let $E$ be a generic search space of exemplars. In our motivating application, $E$ is the space of peptides. Each element $x \in E$ has an unknown binary label $y(x) = \{0, 1\}$. A known deterministic function $f(x)$ measures the cost or disutility associated with $x$. Our goal is to perform experiments so as to find $x$ such that it has positive label and its cost function $f(x)$ is minimum.

To obtain labels of exemplars, we can do a batch of experiments, which evaluate a subset $S \subseteq E$ and obtain labels at each time. we measure quality

of $S$ by

$$f^*(S) = \begin{cases} \min\limits_{x \in S: y(x) = 1} f(x) & \text{if } \{x \in S : y(x) = 1\} \neq \emptyset \\ \infty & \text{if } \{x \in S : y(x) = 1\} = \emptyset \end{cases} \quad (1)$$

Let $b$ be a target value and we wish to find $S \subseteq E$ such that $f^*(S)$ is, in some sense, better than $b$. Specifically, we consider the following two measures:

$$\begin{aligned} \text{Probability of Improvement:} \quad & \text{P}^*(S) = \mathbb{P}(f^*(S) < b) \\ \text{Expected Improvement:} \quad & \text{EI}(S) = \mathbb{E}[(b - f^*(S))] \end{aligned} \quad (2)$$

We wish to find $S$ that maximize one of these two measures. Let $g(S)$ be either $\text{P}^*(S)$ or $\text{EI}(S)$ and let the cardinality of $S$ be the only constraint on $S$. Our goal is then:

$$\max_{S \subseteq E: |S| < k} g(S) \quad (3)$$

## 2.3 Solution Method for EI

If $g(S)$ is $\text{EI}(S)$, from equation (2) (3), our goal becomes

$$\max_{S \subseteq E: |S| \leq k} \mathbb{E}\left[(b - f^*(S))^+\right] \quad (4)$$

We first prove that the objective function is submodular, and then describe our greedy approach to solve (4), finally we show that we have guarantee for our greedy policy compared with optimal policy.

### 2.3.1 Submodularity

### 2.3.2 Greedy Approach

Suppose we have chosen $S = \{x_1, x_2, \ldots, x_n\}$ as a batch of points we are going to evaluate next, and if we want to incorporate one more point $e$, which is distinct from $x_1, x_2, \ldots, x_n$, such that the expected improvement increases most, we use the following criterion to find $e$:

$$\arg\max_{e \in E \backslash S} \mathbb{E}\left[(b - f^*(S \cup \{e\}))^+\right] \quad (5)$$

2

From (1) we write expected improvement part in (3) as

$$\mathbb{E}\left[(b - f^*(S))^+\right] = \mathbb{E}[b - \min_{x \in S \cup \{e\}:y(x)=1} f(x)]$$

$$= \begin{cases} \mathbb{E}[b - f^*(S)] & \text{if } y(e) = 0 , \\ \mathbb{E}[b - \min\{f(e), f^*(S)\}] & \text{if } y(e) = 1 , \end{cases}$$

$$= \mathbb{E}[b - f^*(S) + \mathbb{1}_{\{y(e)=1, f(e)<f^*(S)\}}[f^*(S) - f(e)]]$$

After some algebra we can write (5) as

$$\arg\max_{e \in E \setminus S} \mathbb{E}[\mathbb{1}_{\{y(e)=1, f(e)<f^*(S)\}}[f^*(S) - f(e)]]$$

$$= \arg\max_{e \in E \setminus S} \sum_{i=1}^{|S|} \mathbb{P}(y(e) = 1, y(x_i) = 1, y(x_j) = 0, \forall j < i)[f(x_i) - f(e)]^+$$

$$+ \mathbb{P}(y(e) = 1, y(x_j) = 0, \forall j)[b - f(e)]^+$$

where $f(x_i) <= f(x_j)$ for $\forall i < j, x_i, x_j \in S$. Since

$$\mathbb{P}(y(e) = 1, y(x_i) = 1, y(x_j) = 0, \forall j < i)$$
$$= \mathbb{P}(y(x_1) = 0)\mathbb{P}(y(x_2) = 0|y(x_1) = 0) \dots \mathbb{P}(y(e) = 1|\mathcal{F}(x_1, x_2, \dots, x_i))$$
$$\propto \mathbb{P}(y(e) = 1|\mathcal{F}(x_1, x_2, \dots, x_i))$$