Bayesian Optimization for Peptide Design

Peter Frazier
Associate Professor, Operations Research & Information Engineering, Cornell
Staff Data Scientist, Uber

Lori Tallorin (UCSD)
Jialei Wang (Cornell)
Eunice Kim (UCSD)
Swagut Sahu (UCSD)
Nick Kosa (UCSD)
Pu Yang (Cornell)

Matt Thompson (Northwestern)
Mike Gilson (UCSD)
Nathan Gianneschi (Northwestern)
Mike Burkart (UCSD)

“Discovering de novo peptide substrates for enzymes using machine learning.” Nature Communications, 2018
Pharmaceutical research and development expenditure in the U.S. from 1980 to 2017 (in billion U.S. dollars)

Source: PhRMA
© Statista 2018

Additional Information:
United States; PhRMA members
Goal: accelerate the discovery of biomaterials and drugs

Scientist

I want to find \( x \) such that \( f(x) \) is large!

Measure of how well the experiment worked: \( f(x) \)

Experiments

Decision about what experiment to perform next: \( x \)
BayesOpt can help

Decision about what point to evaluate next:

Objective function: \( f(x) \)

Algorithm

I want to find \( x \) such that \( f(x) \) is large!
We use BayesOpt to design peptides for biomedical applications.
Biology primer: What is a peptide?

• A peptide is a sequence of amino acids. Most of our peptides will be between 5 and 35 amino acids long.

• An amino acid is a molecule. There are 20 of them in nature. We represent them by capital letters.

• A peptide of length 9: DSLEFSKIA
We use BayesOpt to design peptides that are active in this chemical reaction.
This will create more flexible peptide-based "labels"

Existing labels are very useful...

**Green Fluorescent Protein (GFP):**
GKLPVPWPWPTVTTFSYGVQCFSRYPD
HMKQHDFKSAMPEGYIQERTIFFK
DDGNYKSAEVKFEQDTLVNRIELTG
TDFKEDGNILGNKMEYNNAHNVYI
MTDKAKNIKVNFKIRHNEDGSVQ
LADHYQQNPITPDGPVLLPDNHLYLSTQSTLSDPNEKRDHMIYFEFVTAAA
IT

Uses:
- tracking proteins in cells;
- measuring protein quantities;
- drug discovery
This will create more flexible peptide-based "labels"

Existing labels are very useful...

Green Fluorescent Protein (GFP):
GKLPVPWPTLVTTSYGVQCFSRYPD
HMKQHDFFKSAMPEGYIQERTIFFK
DDGNYKSAEVKFEGLNRIELTG
TDFKEDGNILGNKMEYNNAHNVI
MTDKAKNGIKVNFIRHNIEDGSVQ
LADHYQNPIDGPVLLPDNHYS
TQSTLSKDPNEKRDHMIYFEFVTAAA
IT

...But are built into the organism, and are "always on"

Our labels can be turned on with the addition of an enzyme, allowing us to study dynamic cellular processes, and to trigger fluorescence based on a "chemical signal"

Winner of the 2008 Nobel Prize in Chemistry
To be useful, our peptide must be short

Recall:

- For a peptide labeling system to be useful, it should be short (goal: no more than ~12 amino acids)

- Otherwise it may change the behavior of the proteins in which it is embedded
It is hard to find short hits; Math makes it easier.

- Hits are rare: about 1 in $10^5$ among shorter peptides.

- Testing peptides is time-consuming

- We test 500 peptides at time. $500 << 10^5$.

- To help us, we have some known hits, obtained from natural organisms. They are too long to be used directly.
Here’s how we test peptides
Our method is called: **Peptide Optimization with Optimal Learning (POOL)**

POOL has 2 parts:

1. Predict which peptides are “hits”, using a simple interpretable Bayesian classifier

2. Use these predictions in an intelligent way to recommend a set of recommend to test next
Our method is called: Peptide Optimization with Optimal Learning (POOL)

POOL has 2 parts:

1. Predict which peptides are “hits”, using a simple interpretable Bayesian classifier
2. Use these predictions in an intelligent way to recommend a set of recommend to test next
We use (Bayesian) Naive Bayes

- Disadvantages: a more modern method could improve accuracy

- Advantages:
  - Interpretability: easy to explain to collaborators, debug, and customize
  - Strong prior => Robust to extremely small amounts of data
  - Good uncertainty quantification (vs. deep learning)
  - Computational scalability!
Naive Bayes assumes two latent matrices

\( \theta^{(\text{hit})}, \ P(\text{amino acid} \mid \text{hit}) \)  

| Position relative to Serine | -10 | -9  | -8  | -7  | -6  | -5  | -4  | -3  | -2  | -1  | 0  | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
|-----------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|----|----|----|----|----|----|----|----|----|----|----|
| DE                          | .19 | .08 | .07 | .07 | .28 | .45 | .07 | .09 | .05 | .72 | .04 | .37 | .06 | .32 | .07 | .07 | .14 | .07 | .07 | .07 | .14 | .08 |
| NQ                          | .08 | .10 | .15 | .07 | .14 | .07 | .07 | .06 | .05 | .04 | .04 | .05 | .06 | .06 | .06 | .07 | .07 | .07 | .12 | .07 | .07 | .24 |
| FWY                         | .08 | .08 | .27 | .27 | .07 | .07 | .15 | .06 | .08 | .04 | .04 | .05 | .11 | .06 | .06 | .07 | .07 | .07 | .07 | .07 | .07 | .29 | .28 |
| HKR                         | .10 | .08 | .12 | .07 | .10 | .15 | .06 | .06 | .05 | .04 | .04 | .05 | .06 | .06 | .07 | .11 | .07 | .16 | .08 | .14 | .09 | .07 | .08 |
| AILMV                       | .21 | .12 | .10 | .27 | .20 | .07 | .46 | .09 | .61 | .04 | .61 | .21 | .14 | .58 | .19 | .45 | .33 | .34 | .29 | .28 | .14 | .09 | .07 | .08 |
| GP                          | .08 | .08 | .07 | .07 | .07 | .07 | .06 | .54 | .05 | .04 | .04 | .05 | .06 | .06 | .06 | .07 | .14 | .09 | .07 | .07 | .07 | .07 | .07 | .08 |
| ST                          | .19 | .40 | .14 | .10 | .07 | .07 | .06 | .05 | .04 | .15 | .18 | .46 | .06 | .17 | .14 | .14 | .21 | .07 | .10 | | | | |
| C                           | .08 | .07 | .07 | .07 | .07 | .07 | .06 | .05 | .04 | .04 | .05 | .06 | .06 | .07 | .07 | .07 | .07 | .07 | .07 | .07 | .07 | .07 | .07 | .08 |

\( \theta^{(\text{miss})}, \ P(\text{amino acid} \mid \text{miss}) \)  

| Position relative to Serine | -10 | -9  | -8  | -7  | -6  | -5  | -4  | -3  | -2  | -1  | 0  | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
|-----------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|----|----|----|----|----|----|----|----|----|----|----|
| DE                          | .20 | .28 | .13 | .07 | .16 | .34 | .06 | .06 | .05 | .56 | .04 | .12 | .06 | .26 | .07 | .07 | .12 | .07 | .07 | .07 | .07 | .08 | .08 |
| NQ                          | .18 | .08 | .13 | .07 | .16 | .07 | .06 | .05 | .08 | .04 | .19 | .06 | .12 | .07 | .07 | .07 | .12 | .07 | .12 | .07 | .12 | .08 | .08 |
| FWY                         | .08 | .08 | .08 | .27 | .16 | .16 | .13 | .06 | .13 | .08 | .04 | .05 | .12 | .06 | .07 | .07 | .10 | .07 | .10 | .07 | .10 | .10 | .10 | .18 |
| HKR                         | .08 | .13 | .13 | .07 | .16 | .07 | .06 | .05 | .08 | .04 | .12 | .15 | .06 | .18 | .07 | .17 | .07 | .17 | .07 | .17 | .30 | .08 | .36 |
| AILMV                       | .08 | .13 | .15 | .29 | .25 | .07 | .39 | .09 | .43 | .04 | .71 | .30 | .38 | .54 | .12 | .37 | .41 | .30 | .19 | .36 |
| GP                          | .08 | .10 | .08 | .07 | .07 | .10 | .55 | .17 | .04 | .04 | .12 | .06 | .06 | .07 | .06 | .07 | .12 | .07 | .07 | .12 | .07 | .07 | .08 |
| ST                          | .23 | .13 | .24 | .07 | .07 | .07 | .13 | .06 | .05 | .12 | .04 | .05 | .12 | .09 | .12 | .23 | .07 | .17 | .07 | .07 | .08 | .08 | .08 |
| C                           | .08 | .08 | .07 | .07 | .07 | .07 | .06 | .05 | .04 | .04 | .05 | .06 | .06 | .07 | .07 | .07 | .07 | .07 | .07 | .07 | .07 | .07 | .08 |

\[
P(y(x) = 1 \mid x, \theta^{\text{hit}}, \theta^{\text{miss}}) = \frac{P(\text{hit}) \prod_i \theta_i^{(\text{hit})}}{P(\text{hit}) \prod_i \theta_i^{(\text{hit})} + P(\text{miss}) \prod_i \theta_i^{(\text{miss})}}
\]
We put independent Dirichlet priors on columns in these matrices

$$\theta^{(hit)}, \ P(\text{amino acid} \mid \text{hit})$$

<table>
<thead>
<tr>
<th></th>
<th>-10</th>
<th>-9</th>
<th>-8</th>
<th>-7</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>.19</td>
<td>.08</td>
<td>.07</td>
<td>.07</td>
<td>.28</td>
<td>.45</td>
<td>.07</td>
<td>.09</td>
<td>.05</td>
<td>.72</td>
<td>.04</td>
<td>.37</td>
<td>.06</td>
<td>.06</td>
<td>.32</td>
<td>.07</td>
<td>.07</td>
<td>.14</td>
<td>.07</td>
<td>.08</td>
<td></td>
</tr>
<tr>
<td>NQ</td>
<td>.08</td>
<td>.10</td>
<td>.15</td>
<td>.07</td>
<td>.14</td>
<td>.07</td>
<td>.07</td>
<td>.06</td>
<td>.05</td>
<td>.04</td>
<td>.04</td>
<td>.05</td>
<td>.06</td>
<td>.06</td>
<td>.06</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.12</td>
<td>.08</td>
<td></td>
</tr>
<tr>
<td>FWY</td>
<td>.08</td>
<td>.08</td>
<td>.27</td>
<td>.27</td>
<td>.07</td>
<td>.07</td>
<td>.15</td>
<td>.06</td>
<td>.08</td>
<td>.04</td>
<td>.04</td>
<td>.05</td>
<td>.11</td>
<td>.06</td>
<td>.06</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.24</td>
<td></td>
</tr>
<tr>
<td>HKR</td>
<td>.10</td>
<td>.08</td>
<td>.12</td>
<td>.07</td>
<td>.10</td>
<td>.15</td>
<td>.06</td>
<td>.06</td>
<td>.05</td>
<td>.04</td>
<td>.04</td>
<td>.05</td>
<td>.06</td>
<td>.06</td>
<td>.07</td>
<td>.07</td>
<td>.11</td>
<td>.07</td>
<td>.16</td>
<td>.08</td>
<td></td>
</tr>
<tr>
<td>AILMV</td>
<td>.21</td>
<td>.12</td>
<td>.10</td>
<td>.27</td>
<td>.20</td>
<td>.07</td>
<td>.46</td>
<td>.09</td>
<td>.61</td>
<td>.04</td>
<td>.61</td>
<td>.21</td>
<td>.14</td>
<td>.58</td>
<td>.19</td>
<td>.45</td>
<td>.33</td>
<td>.34</td>
<td>.29</td>
<td>.28</td>
<td></td>
</tr>
<tr>
<td>GP</td>
<td>.08</td>
<td>.08</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.06</td>
<td>.54</td>
<td>.05</td>
<td>.04</td>
<td>.04</td>
<td>.05</td>
<td>.06</td>
<td>.06</td>
<td>.07</td>
<td>.14</td>
<td>.09</td>
<td>.07</td>
<td>.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST</td>
<td>.19</td>
<td>.40</td>
<td>.14</td>
<td>.10</td>
<td>.07</td>
<td>.07</td>
<td>.06</td>
<td>.06</td>
<td>.05</td>
<td>.04</td>
<td>.15</td>
<td>.18</td>
<td>.46</td>
<td>.06</td>
<td>.17</td>
<td>.14</td>
<td>.14</td>
<td>.21</td>
<td>.07</td>
<td>.10</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>.08</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.06</td>
<td>.06</td>
<td>.05</td>
<td>.04</td>
<td>.04</td>
<td>.05</td>
<td>.06</td>
<td>.06</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.08</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>-10</th>
<th>-9</th>
<th>-8</th>
<th>-7</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>.20</td>
<td>.28</td>
<td>.13</td>
<td>.07</td>
<td>.16</td>
<td>.34</td>
<td>.06</td>
<td>.06</td>
<td>.05</td>
<td>.56</td>
<td>.04</td>
<td>.12</td>
<td>.06</td>
<td>.06</td>
<td>.26</td>
<td>.07</td>
<td>.07</td>
<td>.12</td>
<td>.07</td>
<td>.08</td>
<td></td>
</tr>
<tr>
<td>NQ</td>
<td>.18</td>
<td>.08</td>
<td>.13</td>
<td>.07</td>
<td>.07</td>
<td>.16</td>
<td>.07</td>
<td>.06</td>
<td>.05</td>
<td>.08</td>
<td>.04</td>
<td>.19</td>
<td>.06</td>
<td>.06</td>
<td>.12</td>
<td>.07</td>
<td>.07</td>
<td>.12</td>
<td>.12</td>
<td>.08</td>
<td></td>
</tr>
<tr>
<td>FWY</td>
<td>.08</td>
<td>.08</td>
<td>.08</td>
<td>.27</td>
<td>.16</td>
<td>.16</td>
<td>.13</td>
<td>.06</td>
<td>.13</td>
<td>.04</td>
<td>.04</td>
<td>.05</td>
<td>.12</td>
<td>.06</td>
<td>.06</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.10</td>
<td>.18</td>
<td></td>
</tr>
<tr>
<td>HKR</td>
<td>.08</td>
<td>.13</td>
<td>.13</td>
<td>.07</td>
<td>.16</td>
<td>.07</td>
<td>.06</td>
<td>.05</td>
<td>.08</td>
<td>.04</td>
<td>.12</td>
<td>.15</td>
<td>.06</td>
<td>.18</td>
<td>.07</td>
<td>.17</td>
<td>.07</td>
<td>.30</td>
<td>.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AILMV</td>
<td>.08</td>
<td>.13</td>
<td>.15</td>
<td>.29</td>
<td>.25</td>
<td>.07</td>
<td>.39</td>
<td>.09</td>
<td>.43</td>
<td>.04</td>
<td>.71</td>
<td>.30</td>
<td>.38</td>
<td>.54</td>
<td>.12</td>
<td>.37</td>
<td>.41</td>
<td>.30</td>
<td>.19</td>
<td>.36</td>
<td></td>
</tr>
<tr>
<td>GP</td>
<td>.08</td>
<td>.10</td>
<td>.08</td>
<td>.07</td>
<td>.07</td>
<td>.10</td>
<td>.55</td>
<td>.17</td>
<td>.04</td>
<td>.04</td>
<td>.12</td>
<td>.06</td>
<td>.06</td>
<td>.07</td>
<td>.06</td>
<td>.07</td>
<td>.12</td>
<td>.07</td>
<td>.07</td>
<td>.08</td>
<td></td>
</tr>
<tr>
<td>ST</td>
<td>.23</td>
<td>.13</td>
<td>.24</td>
<td>.07</td>
<td>.07</td>
<td>.13</td>
<td>.06</td>
<td>.05</td>
<td>.12</td>
<td>.04</td>
<td>.05</td>
<td>.12</td>
<td>.09</td>
<td>.12</td>
<td>.23</td>
<td>.07</td>
<td>.17</td>
<td>.07</td>
<td>.07</td>
<td>.08</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>.08</td>
<td>.08</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.06</td>
<td>.06</td>
<td>.05</td>
<td>.04</td>
<td>.04</td>
<td>.05</td>
<td>.06</td>
<td>.06</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.08</td>
<td></td>
</tr>
</tbody>
</table>

- Prior mean is proportional to the number of AA in the class
- The prior on columns far from the Serine is more concentrated close to the mean
Naive Bayes is ok, but far from perfect

- Graph uses training data from ~300 peptides (most are misses.)
- Rates were estimated via leave-one-out cross-validation.
Given imperfect predictions, what should we test next?

- If predictions were perfect, we could just test the shortest peptide predicted to be a hit.
- Our predictions are not perfect.
- How should we decide what to test next?
One simple strategy is:
- Select those peptides with length < target.
- Rank them by predicted probability of a hit
- Test the top 300.

The tested peptides are very similar. If the first tested peptide is not a hit, the other ones probably aren’t either.
Ranking by probability of a hit does not work well

Plot obtained via a simulation experiment
POOL (peptide optimization with optimal learning) works better.

Plot obtained via a simulation experiment.
Let’s do the experiment that maximizes the probability we reach our goal

- Our goal is to find short hits.

- More specifically, our goal is:
  - Find at least one hit of length $b$ or shorter

- Let’s run an experiment that maximizes the probability of reaching this goal.
The best experiment is the solution to a combinatorial optimization problem

- This can be formulated as this combinatorial optimization problem:

\[
\max_{S \subseteq E : |S| \leq k} P(\text{at least one short hit in } S)
\]

- Notation:
  - \( E \) is the set of all peptides.
  - \( S \) is the set of peptides to test.
  - \( k \) is the number of peptides we can test in one experiment. Typically, \( k \) is between 200 and 500.
  - A “short hit” is a hit whose length is less than \( b \).
We can’t solve this exactly, so we approximate its solution using a greedy algorithm

- This combinatorial optimization problem is very challenging: The number of size-k sets of length b peptides is $20^b \binom{b}{k}$. If $b=14$ and $k=500$, this is $10^{19} \binom{14}{500}$.

- Instead, we build up the set $S$ of peptides to test in stages.

- In each stage, find one peptide $e$ to add to $S$ that maximizes the probability of reaching our goal:

$$\max_{e \in E \setminus S} P(\text{at least one short hit in } S \cup \{e\})$$

- Add $e$ to $S$ and repeat, until $S$ has $k=500$ peptides.
The greedy algorithm performs within 63% of optimal

Let $P^*(S) = P(\text{at least one short hit in } S)$.

Lemma: $P^*(S)$ is a monotone submodular functions of $S$.

Proposition: Let $\text{OPT} = \max_{S \subseteq E : |S| \leq k} P^*(S)$, and let GREEDY be the value of the solution obtained by the greedy algorithm. Then

$$\frac{\text{OPT} - \text{GREEDY}}{\text{OPT}} \leq 1 - \frac{1}{e}$$
We can implement the greedy algorithm efficiently

• The greedy optimization step is equivalent to

$$\arg \max_{e \in E \setminus S} P(y(e) = 1 | y(x) = 0 \ \forall x \in S)$$

• We can compute this probability by treating all peptides in S as misses, and re-training our model

• Naive Bayes allows solving the above optimization problem separately for each position in the peptide, making it fast to solve
Here is the intuition why this approach works better than “rank by prob. hit”

- Finding the single peptide to add that maximizes the probability of reaching our goal:
  \[
  \max_{e \in E \setminus S} P(\text{at least one short hit in } S \cup \{e\})
  \]

- Is equivalent to:
  \[
  \max_{e \in E \setminus S} P(e \text{ is a short hit|no short hits in } S)
  \]

- Compare this to the “rank by prob. hit” approach
  \[
  \max_{e \in E \setminus S} P(e \text{ is a short hit})
  \]
POOL works better because its peptides are more diverse

- Peptides added using the value of information approach tend to be **different** from those already in S.

\[
\text{POOL} \quad \max_{e \in E \setminus S} P(e \text{ is a short hit} | \text{no short hits in } S)
\]

\[
\text{ Ranking by prob. of a hit} \quad \max_{e \in E \setminus S} P(e \text{ is a short hit})
\]

- Its recommendations are more **diverse**.
POOL’s recommendations are more diverse
We used POOL to find two peptide labeling systems that can be used simultaneously without cross-talk.
We used POOL to find two peptide labeling systems that can be used simultaneously without cross-talk.
Summary

• POOL (peptide optimization with optimal learning) uses a BayesOpt-style approach to find short peptides with biological activity

• POOL construct a batch of peptides to test by iteratively adding the one that is most likely to succeed, if all others in the batch fail

• POOL works in practice

• Lots of opportunities & grant support in BayesOpt + materials/medicine

Appendix
Sfp-type peptides can also be selectively labeled off-membrane, conjugated to Green Fluorescent Protein (GFP)

We believe we were unable to label our AcpS-type GFP-peptides because of endogeneous AcpS in *E. coli* used to make them
Sequence alignment of hit peptides relative to native PPTase substrates. The sequence alignment for *B. subtilis* PCP, *S. coelicolor* ACP, *E. coli* ACP, YbbR peptide, compared to Sfp-type peptide hits (4P28, 4N28, 4F01) and AcpS-type peptide hits (1F01, 1I04, 3K17, 4T25) to the secondary (α2) structure of *B. subtilis* PCP (PDB: 4MRT). The blue box and red residues show general conserved sequences across all the peptides. The majority of AcpS-type peptides have conserved polar residues in position 2 and 5 (highlighted in yellow). The peptide identification corresponds to the round number and location it was identified from (round number_letter row_spot number on membrane) during the iterative rounds of POOL.
Using VOI to optimize \( P(\geq 1 \text{ short hit}) \) has a shortcoming

- Under our Naïve Bayes model, it is usually possible to increase \( P(\text{hit}) \) by increasing the peptide’s length.

- Thus, the experiments that maximize \( P(\geq 1 \text{ short hit}) \) tend to have length \( b-1 \).

- However, a hit strictly shorter than \( b-1 \) would be even better.

- To allow us to find such strictly shorter peptides, we might consider an alternate goal: expected improvement.
Optimizing expected improvement would fix this

- Let $f(x)$ be the length of peptide $x$.

- $f^*(S) = \min_{x \in S: y(x) = 1} f(x)$ is the length of the shortest hit found.

- Define the **expected improvement** for testing $S$ as:
  $$EI(S) = E[(b - f^*(S))^+]$$

- An $S$ that maximizes $EI(S)$ could contain peptides shorter than $b-1$. 
Efficiently optimizing expected improvement is ongoing work

- Solving $\max_{S \subseteq E: |S| \leq k} \ EI(S)$ exactly is very challenging.

- $EI(S)$ is also a monotone submodular function, and so the greedy algorithm also has an approximation guarantee.

- However, actually finding the single peptide to add that maximizes the expected improvement is itself extremely difficult.

- We are currently using an integer program to do this, but results are pending.
We are greedily optimizing $P(\geq 1 \text{ short hit})$ with one tweak to make real recommendations.

- We have used the following approach in recommending experiments to our collaborators.
  
  - We pre-select a random sequence of lengths $a^1, \ldots, a^k$ strictly less than $b$, and require that the $n^{th}$ peptide selected has length less than $a^n$.

- We then apply the greedy probability of improvement algorithm.

- This improves expected improvement, without hurting $P(\geq 1 \text{ short hit})$.

Expected improvement as a function of $|S|$, estimated via Monte Carlo.
Example:
Recommender Systems
Step 1: Use ML to predict books Jack might enjoy reading

- **The Boy Who Fell from the Sky** (40%)
  - By Jule Owen
  - Kindle Edition: $0.00
  - Paperback: $11.99

- **The Watchers of Eden** (38%)
  - By T.C. Edge
  - Kindle Edition: $0.00
  - Paperback: $13.99

- **Skip** (35%)
  - By Perrin Briar
  - Kindle Edition: $0.00

- **Book of the Night** (30%)
  - By Oliver Pötzsch and Lee Chadeayne
  - Kindle Edition: $0.00
What happens if we use the simple strategy of going with the top 3 most likely to be enjoyed?
What happens if we use the simple strategy of going with the top 3 most likely to be enjoyed?

- Probability he’ll like this book, if he doesn’t like the first one: 40%
- Probability he’ll like this book, if he doesn’t like the first two: 10%
- Probability he’ll like this book, if he doesn’t like the first three: 5%
The probability that Jack likes at least one of these books is only $1 - 0.6 \times 0.9 \times 0.95 = 48.7\%$.
Step 2: Take ML's most recommended book

Probability he'll like this book: 40%
Step 3: Retrain assuming he doesn’t like it

Probability he’ll like this book: 40%
Probability he’ll like this book, if he doesn’t like the first one: 50%
Probability he’ll like this book: 42%
Probability he’ll like this book: 37%
Step 4: Take ML’s most recommended book

- Probability he’ll like this book: 40%
- Probability he’ll like this book, if he doesn’t like the first one: 50%
We are already up to a probability of 70% he’ll like one of these books.

<table>
<thead>
<tr>
<th>Book title</th>
<th>Probability of liking</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Boy Who Fell from the Sky</td>
<td>40%</td>
</tr>
<tr>
<td>Heart of a Champion</td>
<td>50%</td>
</tr>
</tbody>
</table>

Probability he’ll like this book

Probability he’ll like this book, if he doesn’t like the first one
Step 5: Retrain assuming he doesn’t like any of the previously selected books. Take the best one.

Probability he’ll like this book: 40%

Probability he’ll like this book, if he doesn’t like the first one: 50%

Probability he’ll like this book, if he doesn’t like the first two: 20%

Probability he’ll like this book, if he doesn’t like the first three: 15%
Step 5: Retrain assuming he doesn’t like any of the previously selected books. Take the best one.

- Probability he’ll like this book: 40%
- Probability he’ll like this book, if he doesn’t like the first one: 50%
- Probability he’ll like this book, if he doesn’t like the first two: 20%
Providing a diverse selection increases the chance he’ll like at least one from 48.7% to \(1 - 0.6 \times 0.5 \times 0.8 = 76\%\)

<table>
<thead>
<tr>
<th>Probability he’ll like this book</th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability he’ll like this book, if he doesn’t like the first one</td>
<td>50%</td>
</tr>
<tr>
<td>Probability he’ll like this book, if he doesn’t like the first two</td>
<td>20%</td>
</tr>
</tbody>
</table>
We do not try to make every pick a winner

• We didn’t design the selection so that he would like every book selected.

• We designed it so that he would like at least one.

• The last book may be unlikely to be selected. It is designed as a good backup, not a good first pick.