Bayesian Optimization for Materials Design and Drug Discovery

Peter Frazier

Associate Professor, Operations Research & Information Engineering, Cornell

Staff Data Scientist, Uber

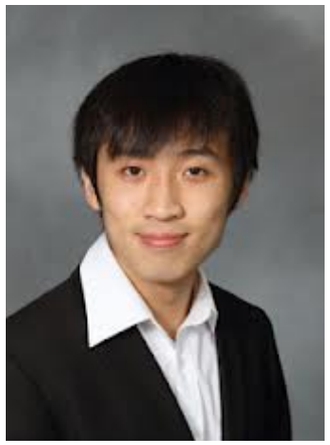
Lori Tallorin
(UCSD)

Jialei Wang
(Cornell)

Eunice Kim
(UCSD)

Swagut Sahu
(UCSD)

Nick Kosa
(UCSD)



Pu Yang
(Cornell)

Matt Thompson
(UCSD/Northwestern)

Mike Gilson
(UCSD)

Nathan Gianneschi
(UCSD/Northwestern)

Mike Burkart
(UCSD)

Science & technology relies on trial & error

- Edison famously used trial & error:
  - ~100 materials tested developing the carbon microphone.
  - ~6000 filaments tested developing a better incandescent bulb.

- In drug discovery, robots often screen > 100,000 compounds.

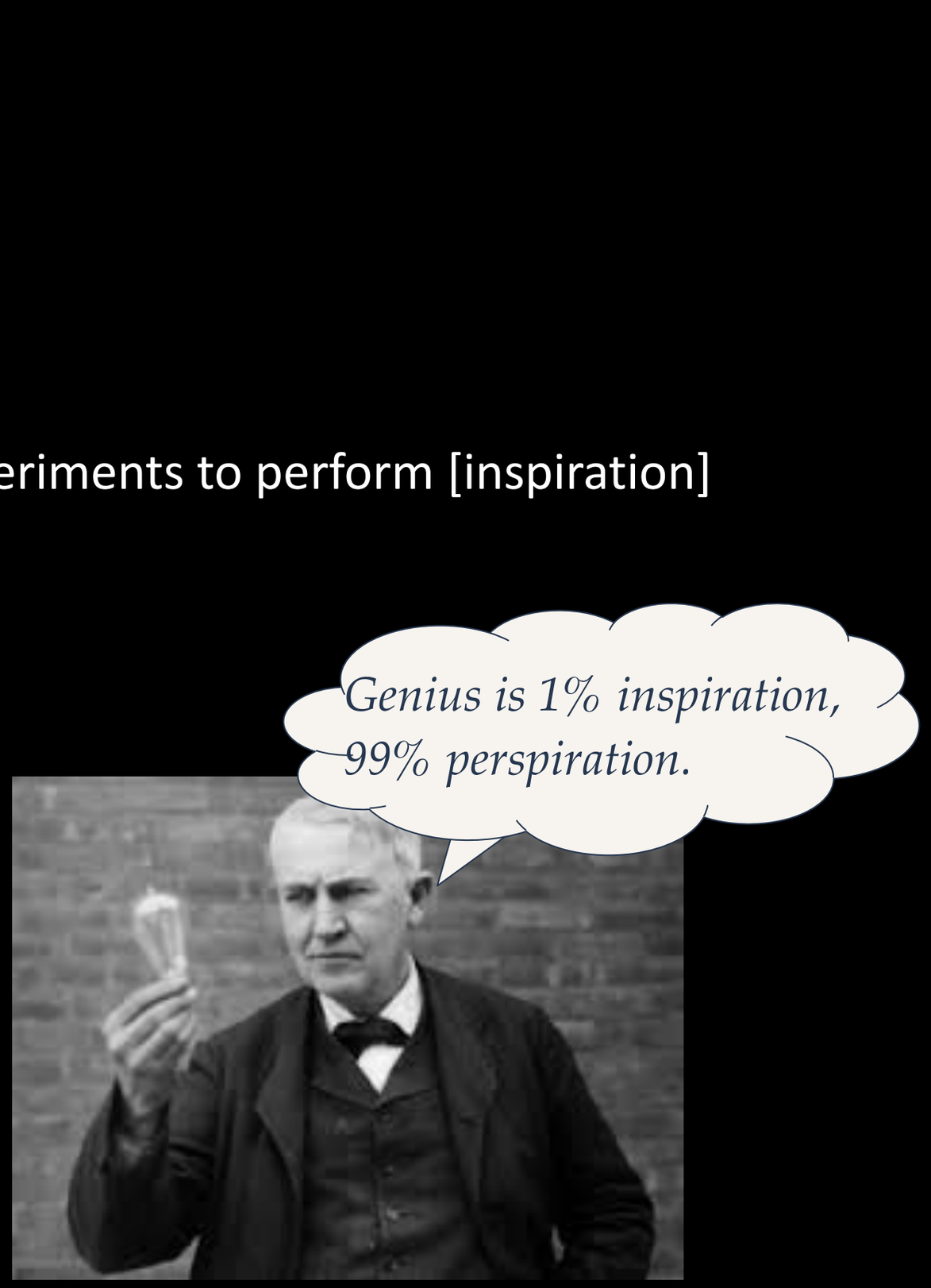
- New polymers and other biomaterials are often created using trial & error.
  [Typically 100s of materials tried.]

Doing trial & error well is really important

- Many good projects fail because trial & error didn’t pan out
- If we could improve trial & error:
  - More technology development efforts would succeed
  - Scientists could take on more ambitious projects
  - Technological development would get faster

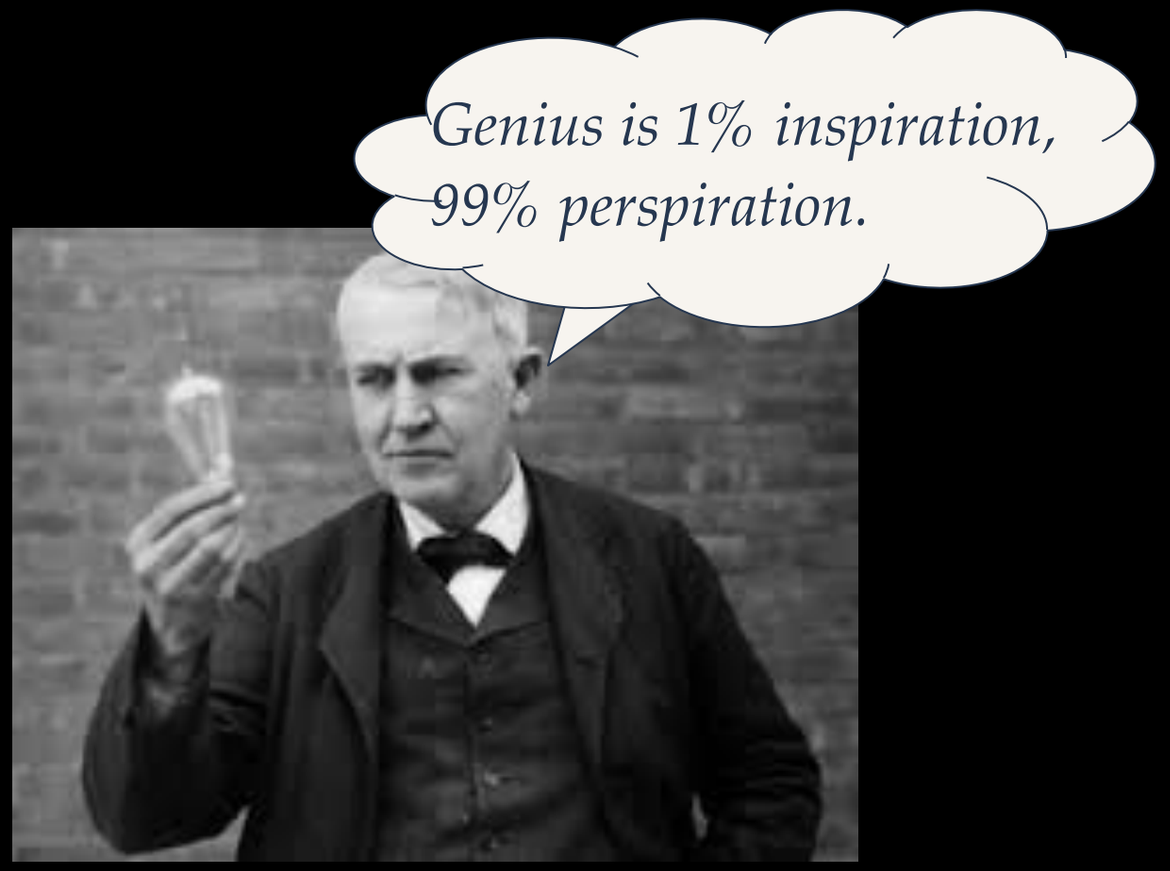
We want to improve trial & error

- Keys to doing trial & error well:
  - Make intelligent decisions about experiments to perform [inspiration]
  - Do lots of experiments [perspiration]
  - Be lucky



We want to improve trial & error

- Keys to doing trial & error well:
  - **Make intelligent decisions about experiments to perform [inspiration]**
  - Do lots of experiments [perspiration]
  - Be lucky

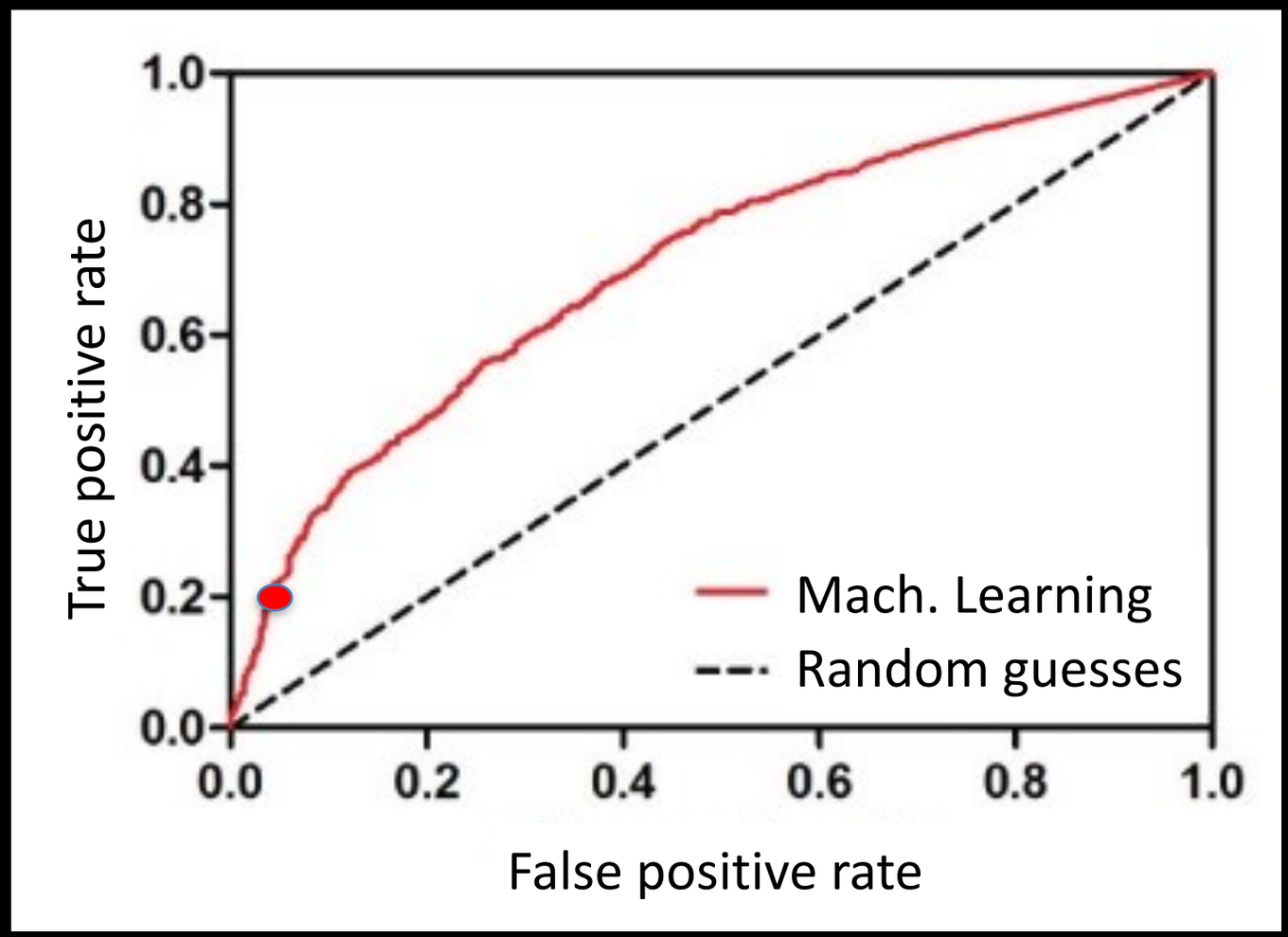


Machine learning is often seen as a magic box that predicts things

Here's how machine learning is typically used in materials discovery

1. Feed in peptides (or small molecule, polymers, alloys, …) synthesized in the past with property measurements ("Training Data")
2. Machine learning predicts properties for unsynthesized peptides
3. Rank peptides by predicted desirability
4. Test the top 10

Challenge:
Machine learning makes errors



The machine learning algorithm with a threshold corresponding to the red dot will label:

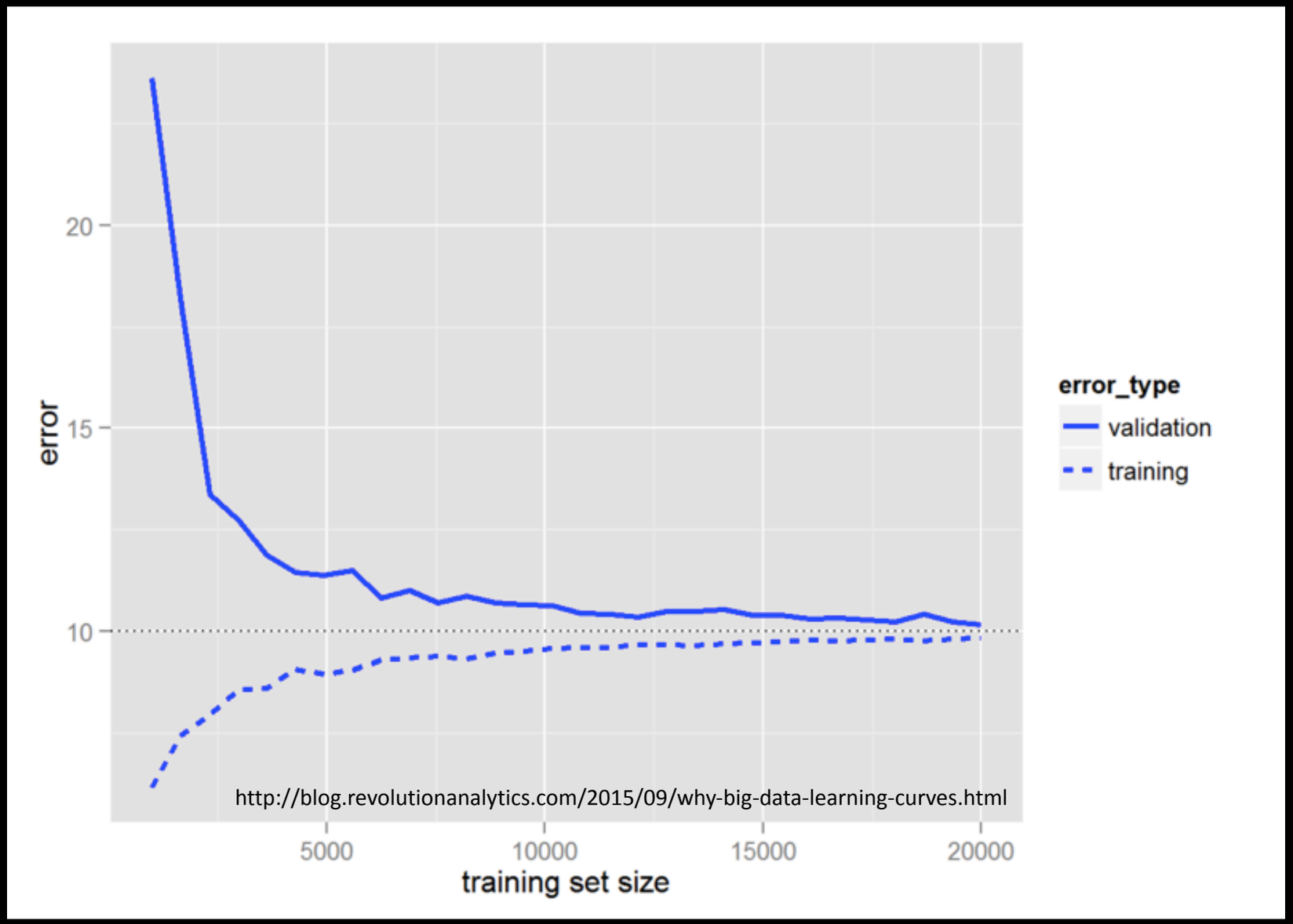
5% of non-active peptides as active [ideally, this would be 0%]

20% of active peptides as active [ideally, this would be 100%]

If 1 in 10^5 peptides are active,

we'll have to test >10^4 predicted-active peptides to find the first active one.

Challenge: Machine learning makes BIG errors when it only has a little training data



In materials discovery, we usually have only a little training data for the property of interest. [Often < 10, < 100]

Solutions

- Include physical knowledge into the machine learning model to make it more accurate
- Get more training data, if you can
- Build fancier machine learning methods [usually requires more training data]
- Use machine learning in an intelligent way

Solutions

- Include physical knowledge into the machine learning model to make it more accurate
- Get more training data, if you can
- Build fancier machine learning methods [usually requires more training data]
- Use machine learning in an intelligent way

Example:

Recommendation Systems

Step 1: Use machine learning to predict books Jack might enjoy reading

40%

The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

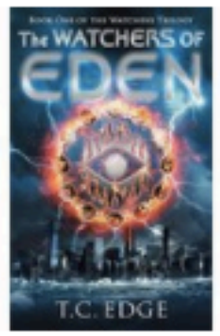
Kindle Edition
$0.00
Whispersync for Voice-ready

Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices
$6.65 used & new (33 offers)

Other Formats: Audible Audio Edition

38%



The Watchers of Eden: The Watchers Trilogy (The Watchers Series Book 1)
by T.C. Edge

Kindle Edition
$0.00
Auto-delivered wirelessly

Paperback
$13.99 Prime
Get it by **Saturday, Dec 10**

More Buying Choices
$13.99 used & new (3 offers)

35%

Skip: An Epic Science Fiction Fantasy Adventure Series (Book 1) Dec 18, 2014
by Perrin Briar

Kindle Edition
$0.00
Auto-delivered wirelessly

30%

Best Seller

Book of the Night: The Black Musketeers Oct 4, 2016
by Oliver Pötzsch and Lee Chadeayne

Kindle Edition
$0.00 kindleunlimited
Read this and over 1 million books with Kindle Unlimited.

$3.99 to buy
Whispersync for Voice-ready

What happens if we use the simple strategy of going with the top 3 most likely to be enjoyed?

40%

The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

Kindle Edition
$0.00
Whispersync for Voice-ready

Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices
$6.65 used & new (33 offers)

Other Formats: Audible Audio Edition

38%

The Watchers of Eden: The Watchers Trilogy (The Watchers Series Book 1)
by T.C. Edge

Kindle Edition
$0.00
Auto-delivered wirelessly

Paperback
$13.99 Prime
Get it by **Saturday, Dec 10**

More Buying Choices
$13.99 used & new (3 offers)

35%

Skip: An Epic Science Fiction Fantasy Adventure Series (Book 1) Dec 18, 2014
by Perrin Briar

Kindle Edition
$0.00
Auto-delivered wirelessly

30%

Best Seller

Book of the Night: The Black Musketeers Oct 4, 2016
by Oliver Pötzsch and Lee Chadeayne

Kindle Edition
$0.00 kindleunlimited
Read this and over 1 million books with Kindle Unlimited.

$3.99 to buy
Whispersync for Voice-ready

What happens if we use the simple strategy of going with the top 3 most likely to be enjoyed?

40%

The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

Kindle Edition
$0.00
Whispersync for Voice-ready

Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices
$6.65 used & new (33 offers)

Other Formats: Audible Audio Edition

Probability he'll like this book, if he doesn't like the first one

10%

The Watchers of Eden: The Watchers Trilogy (The Watchers Series Book 1)
by T.C. Edge

Kindle Edition
$0.00
Auto-delivered wirelessly

Paperback
$13.99 Prime
Get it by **Saturday, Dec 10**

More Buying Choices
$13.99 used & new (3 offers)

Probability he'll like this book, if he doesn't either of the first two

5%

Skip: An Epic Science Fiction Fantasy Adventure Series (Book 1) Dec 18, 2014
by Perrin Briar

Kindle Edition
$0.00
Auto-delivered wirelessly

Best Seller

Book of the Night: The Black Musketeers Oct 4, 2016
by Oliver Pötzsch and Lee Chadeayne

Kindle Edition
$0.00 kindleunlimited
Read this and over 1 million books with Kindle Unlimited.

$3.99 to buy
Whispersync for Voice-ready

The probability that Jack likes at least one of these books is only

1-.6*.9*.95= 48.7%

40%

The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

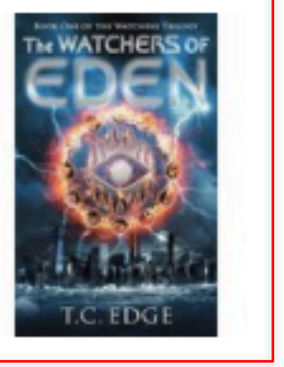
Kindle Edition
$0.00
Whispersync for Voice-ready

Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices
$6.65 used & new (33 offers)

Other Formats: Audible Audio Edition

10%



The Watchers of Eden: The Watchers Trilogy (The Watchers Series Book 1)
by T.C. Edge

Kindle Edition
$0.00
Auto-delivered wirelessly

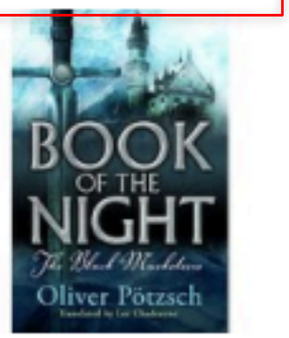
Paperback
$13.99 Prime
Get it by **Saturday, Dec 10**

More Buying Choices
$13.99 used & new (3 offers)

5%

Skip: An Epic Science Fiction Fantasy Adventure Series (Book 1) Dec 18, 2014
by Perrin Briar

Kindle Edition
$0.00
Auto-delivered wirelessly



Best Seller

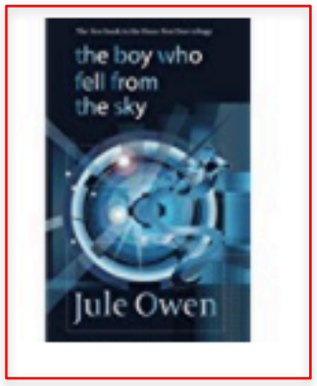
Book of the Night: The Black Musketeers Oct 4, 2016
by Oliver Pötzsch and Lee Chadeayne

Kindle Edition
$0.00 kindleunlimited
Read this and over 1 million books with Kindle Unlimited.

$3.99 to buy
Whispersync for Voice-ready

Step 2: Take machine learning's most recommended book.

Probability he'll like this book 40%



The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

Kindle Edition
$0.00
Whispersync for Voice-ready

Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices
$6.65 used & new (33 offers)

Other Formats: Audible Audio Edition

Step 3: Retrain assuming he doesn't like it

Probability he'll like this book 40%

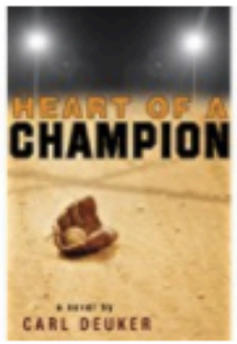
The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

Kindle Edition
$0.00
Whispersync for Voice-ready

Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices

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



Heart of a Champion May 30, 2009
by Carl Deuker

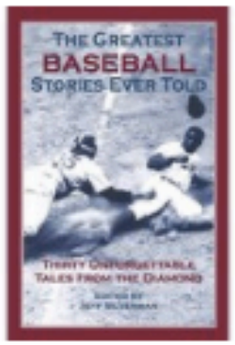
Kindle Edition
$9.99
Auto-delivered wirelessly

Paperback
$6.35 ~~$9.99~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices
$0.01 used & new (187 offers)

Other Formats: Mass Market Paperback, Library Binding

42%



The Greatest Baseball Stories Ever Told
by Jeff Silverman

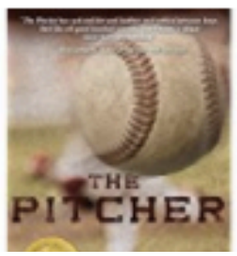
Paperback
$13.12 ~~$16.95~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices
$3.50 used & new (96 offers)

Hardcover
$15.61 used & new (24 offers)

Other Formats: Audio CD

37%



The Pitcher Oct 3, 2016
by William Hazelgrove

Kindle Edition
$6.29
Auto-delivered wirelessly

Step 4: Take machine learning's most recommended book

Probability he'll like this book 40%

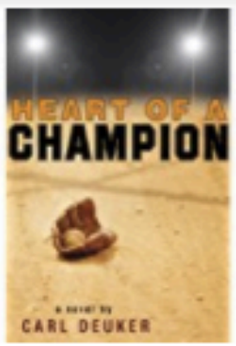
The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

Kindle Edition
$0.00
Whispersync for Voice-ready

Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices

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



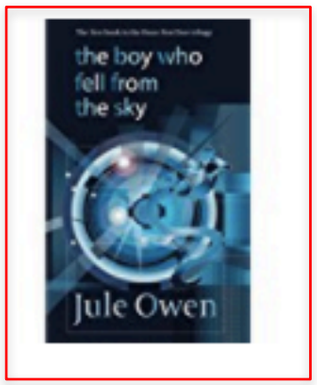
Heart of a Champion May 30, 2009
by Carl Deuker

Kindle Edition
$9.99
Auto-delivered wirelessly

Paperback
$6.35 ~~$9.99~~ Prime
Get it by **Saturday, Dec 10**

We are already up to a probability of 70% he'll like one of these books

Probability he'll like this book 40%



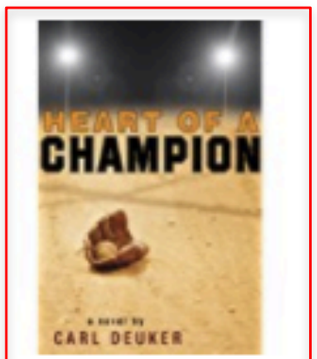
The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

Kindle Edition
$0.00
Whispersync for Voice-ready

Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices

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



Heart of a Champion May 30, 2009
by Carl Deuker

Kindle Edition
$9.99
Auto-delivered wirelessly

Paperback
$6.35 ~~$9.99~~ Prime
Get it by **Saturday, Dec 10**

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%

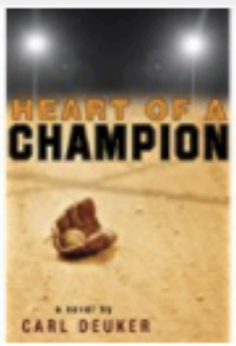
The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

Kindle Edition
$0.00
Whispersync for Voice-ready

Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices

Probability he'll like this book, if he doesn't like the first one — 50%

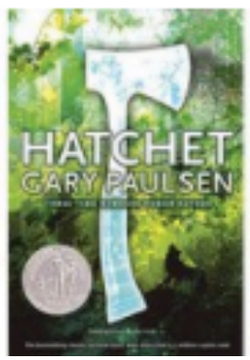


Heart of a Champion May 30, 2009
by Carl Deuker

Kindle Edition
$9.99
Auto-delivered wirelessly

Paperback
$6.35 ~~$9.99~~ Prime
Get it by **Saturday, Dec 10**

Probability he'll like this book, if he doesn't like either of the first two — 20%



Hatchet (Brian's Saga Book 1) Aug 25, 2009
by Gary Paulsen

Kindle Edition
$6.99
Whispersync for Voice-ready

Paperback
$6.00 ~~$7.99~~ Prime
Get it by **Saturday, Dec 10**

15%

Best Seller

The Boys Who Challenged Hitler: Knud Pedersen and the Churchill Clu
by Phillip Hoose

Kindle Edition
$7.99
Auto-delivered wirelessly

Hardcover

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%

The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

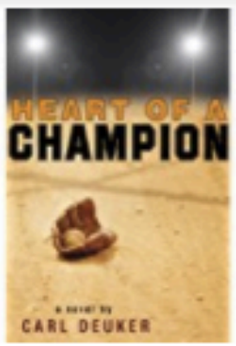
Kindle Edition
$0.00
Whispersync for Voice-ready

Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices

Probability he'll like this book, if he doesn't like the first one

50%



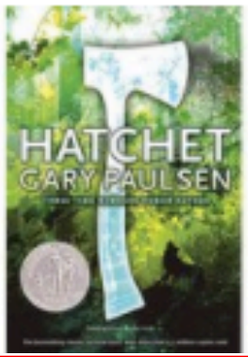
Heart of a Champion May 30, 2009
by Carl Deuker

Kindle Edition
$9.99
Auto-delivered wirelessly

Paperback
$6.35 ~~$9.99~~ Prime
Get it by **Saturday, Dec 10**

Probability he'll like this book, if he doesn't like either of the first two

20%



Hatchet (Brian's Saga Book 1) Aug 25, 2009
by Gary Paulsen

Kindle Edition
$6.99
Whispersync for Voice-ready

Paperback
$6.00 ~~$7.99~~ Prime
Get it by **Saturday, Dec 10**

By providing a diverse selection of books, the chance he'll like at least one is 1-.6*.5*.8=76% (> 45%)

Probability he'll like this book

40%

The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

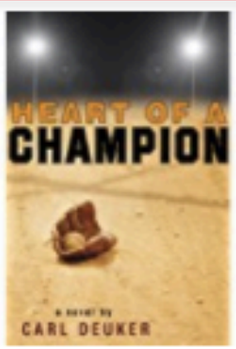
Kindle Edition
$0.00
Whispersync for Voice-ready

Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

More Buying Choices

Probability he'll like this book, if he doesn't like the first one

50%



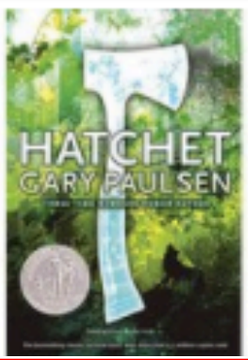
Heart of a Champion May 30, 2009
by Carl Deuker

Kindle Edition
$9.99
Auto-delivered wirelessly

Paperback
$6.35 ~~$9.99~~ Prime
Get it by **Saturday, Dec 10**

Probability he'll like this book, if he doesn't like either of the first two

20%



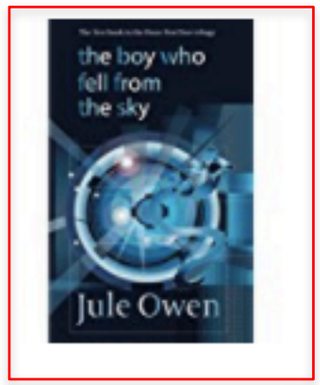
Hatchet (Brian's Saga Book 1) Aug 25, 2009
by Gary Paulsen

Kindle Edition
$6.99
Whispersync for Voice-ready

Paperback
$6.00 ~~$7.99~~ Prime
Get it by **Saturday, Dec 10**

We does not try to make every pick a winner

- We didn't design the selection so that he would like <u>every</u> book selected.
- We designed it so that he would like <u>at least one</u>.
- The last book may be unlikely to be selected. It is designed as a good backup, not a good first pick.

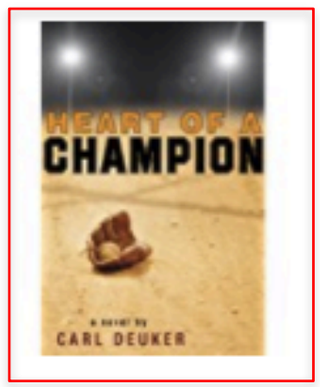


The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

Kindle Edition
$0.00
Whispersync for Voice-ready

Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

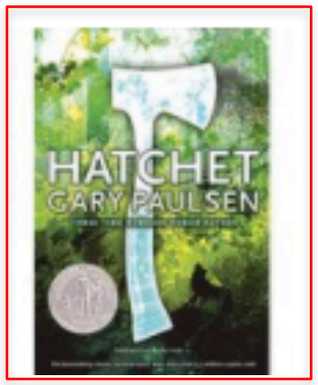
More Buying Choices



Heart of a Champion May 30, 2009
by Carl Deuker

Kindle Edition
$9.99
Auto-delivered wirelessly

Paperback
$6.35 ~~$9.99~~ Prime
Get it by **Saturday, Dec 10**



Hatchet (Brian's Saga Book 1) Aug 25, 2009
by Gary Paulsen

Kindle Edition
$6.99
Whispersync for Voice-ready

Paperback
$6.00 ~~$7.99~~ Prime
Get it by **Saturday, Dec 10**

These ideas come from the literature on Bayesian optimization

Bayesian optimization [Kushner 1964; Mockus 1989; Jones, Schonlau, Welch 1998] is a kind of sequential Bayesian experimental design, for optimizing expensive-to-evaluate functions

In Bayesian optimization, we:

1. Build a Bayesian machine learning model of the objective, based on training data
2. Suggest experiments to run with the largest value of information [Howard 1966]

Bayesian optimization is in a larger class of “optimal learning” methods

Bayesian optimization isn't the only way to improve trial & error in chemistry

- Classical experimental design & response surface methods
- Chemoinformatics &
  quantitative structure activity relationship (QSAR) modeling
- Simulations of chemical systems
- Combinatorial chemistry
- Robotics for high-throughput screening

We have been developing better ways to use machine learning in materials and drug discovery in these problems

- Developing orthogonal protein labels
- Drug development for Ewing's sarcoma
- Finding peptides that bind specifically to metals
- Developing organic semiconductors

We have been developing better ways to use machine learning in materials and drug discovery in these problems

- **Developing orthogonal protein labels.**
- Drug development for Ewing's sarcoma.
- Finding peptides that bind specifically to metals.
- Developing organic semiconductors.

We are using Bayesian optimization to develop orthogonal protein labels

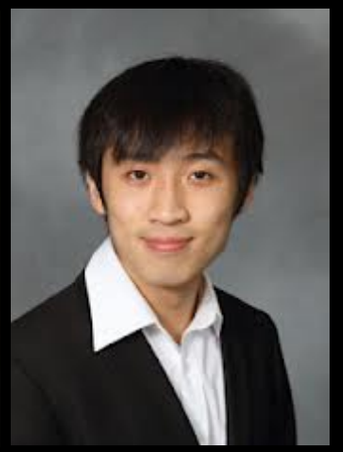
Lori Tallorin (UCSD)

Jialei Wang (Cornell)

Eunice Kim (UCSD)

Swagut Sahu (UCSD)

Nick Kosa (UCSD)



Pu Yang (Cornell)

Matt Thompson (UCSD/ Northwestern)

Mike Gilson (UCSD)

Nathan Gianneschi (UCSD/ Northwestern)

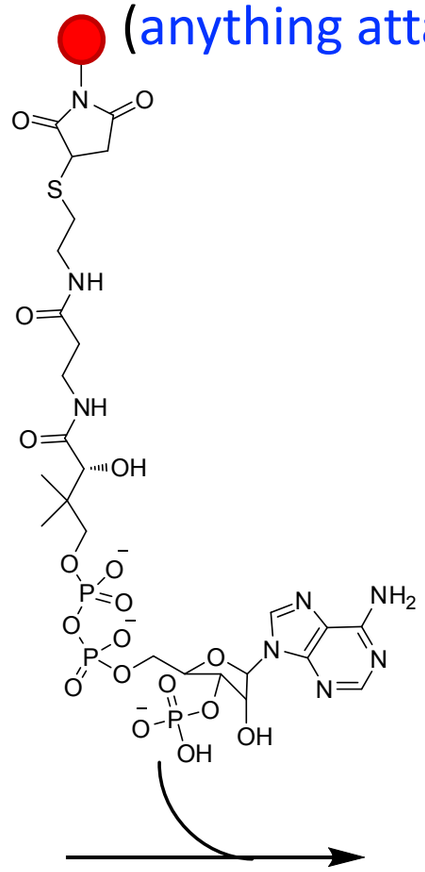
Mike Burkart (UCSD)

Our goal is to build a way to stick things to proteins

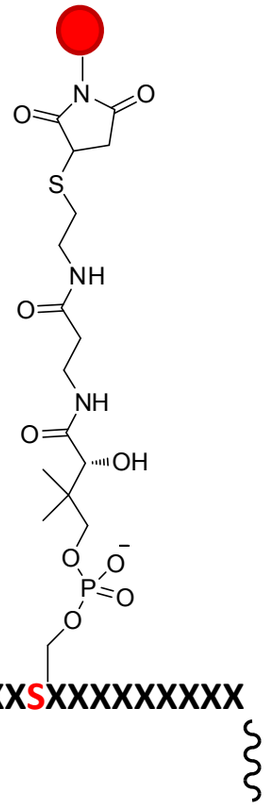
Our goal is to build a way to stick things to proteins

arbitrary label, e.g., red dye
(anything attached to pin)

Phosphopantetheine
(pin)



PPTase enzyme
(person sticking pin into pincushion)



XXXSXXXXXXXXX

OH

XXXSXXXXXXXXX

Peptide substrate
(pincushion),
can be embedded in large protein
(person wearing the pincushion)

Our goal is to build two orthogonal ways to stick things to proteins

- We work with 2 different PPTase enzymes: Sfp and AcpS
- We want to find:

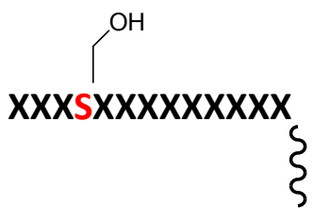
  (1) an Sfp-specific peptide substrate labeled by Sfp but not by AcpS

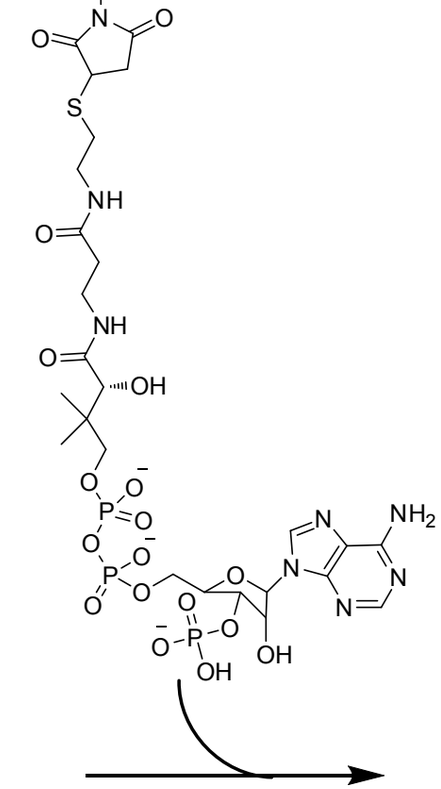
  (2) an AcpS-specific peptide substrate labeled by AcpS but not by Sfp

arbitrary label, e.g., red dye
(anything attached to pin)

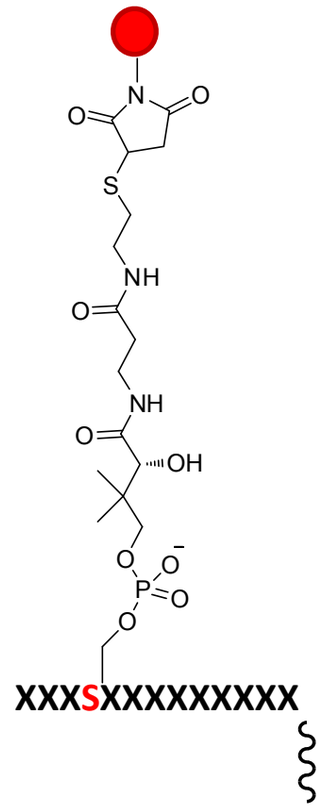
Phosphopantetheine
(pin)

Peptide
substrate
(pincushion),
can be embedded in large protein
(person wearing the pincushion)

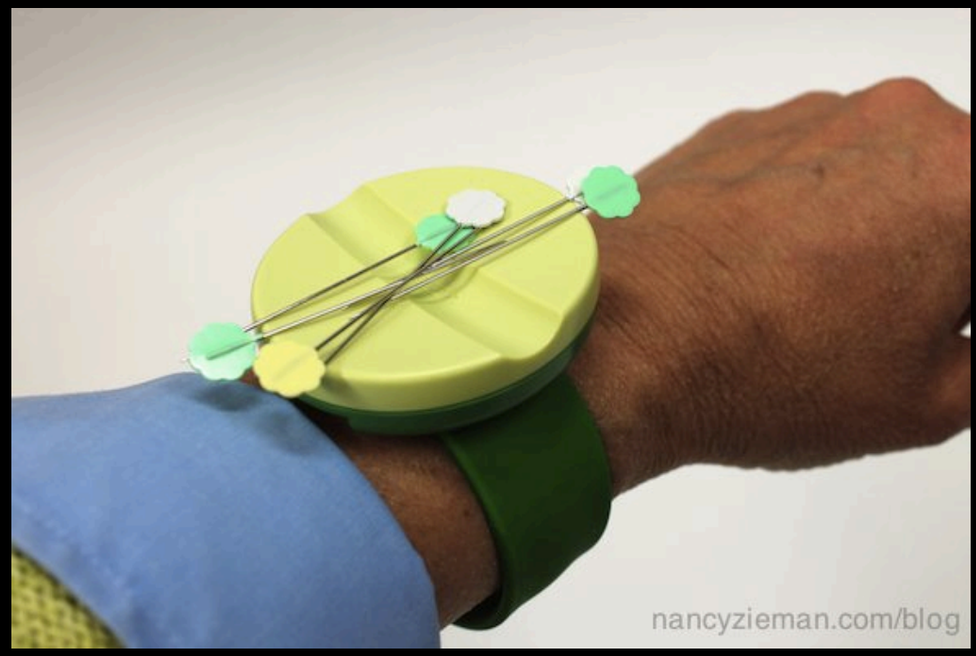




PPTase enzyme
(person sticking pin into pincushion)



Our goal is to build two orthogonal ways to stick things to proteins



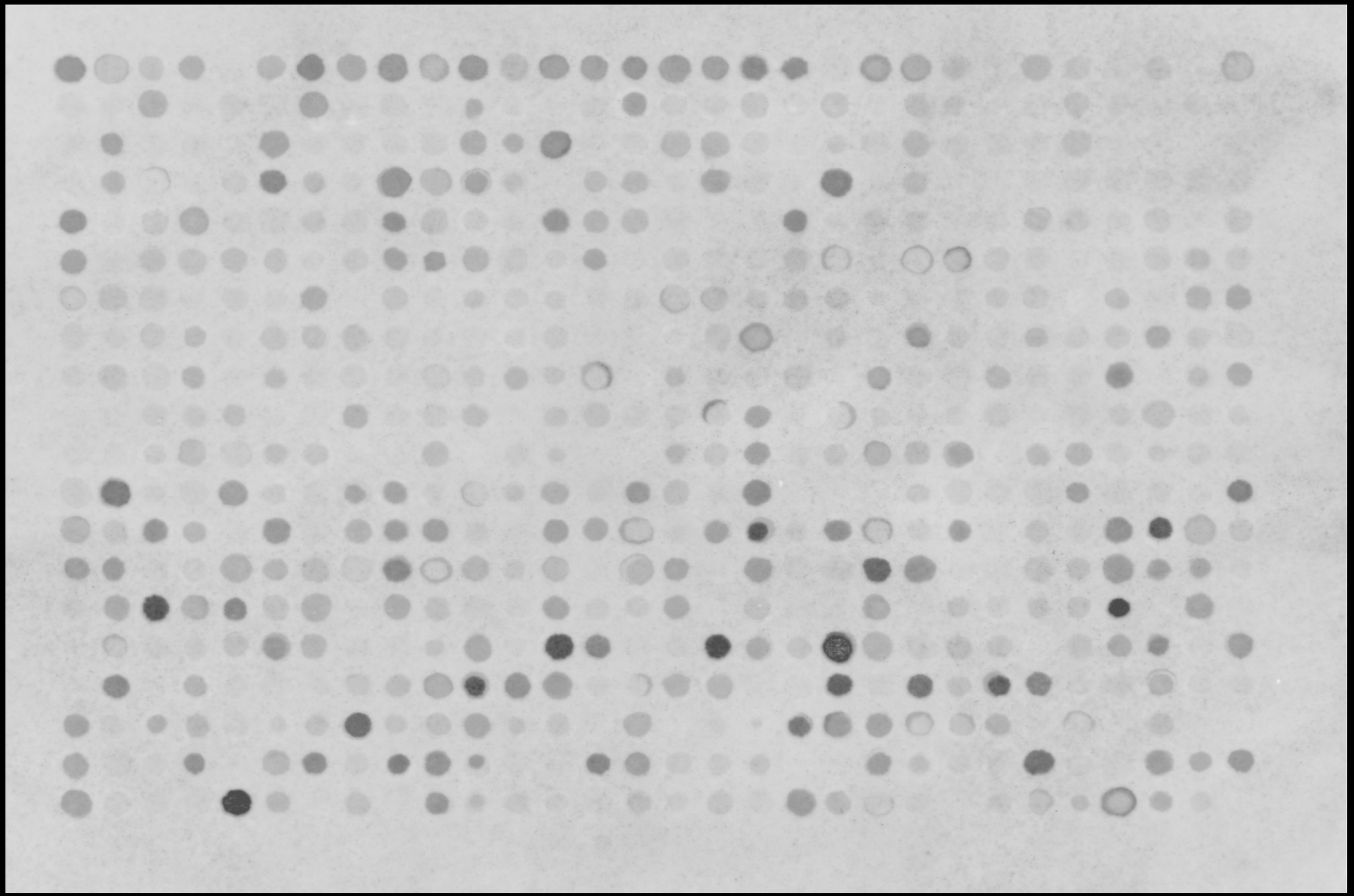
To make our orthogonal labeling system useful, we need the substrates to be short

- If a peptide is a substrate for Sfp and not AcpS, we say it is an “Sfp-specific hit”
- AcpS-specific hits are defined similarly
- For the orthogonal labeling system to be useful, the peptide should be short (say, 8-12 amino acids)
- Otherwise they will change the behavior of the proteins where they are embedded

It is hard to find short hits; Math makes it easier.

- If a peptide allows both chemical reactions to occur, we call it a “hit”.
- Hits are rare: about 1 in 10^5 among shorter peptides.
- Testing peptides is expensive & time-consuming:
  ~1 week from an experimentalist & expensive capacity-limited machine; + material costs
- We test 500 peptides at time. 500 is much smaller than 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



We reduce the experimental effort required to find minimal substrates

- We provide a method for
  Peptide Optimization with Optimal Learning (**POOL**)

- POOL has 2 parts:

1. Predict which peptides are “hits”,
   using a simple machine learning method

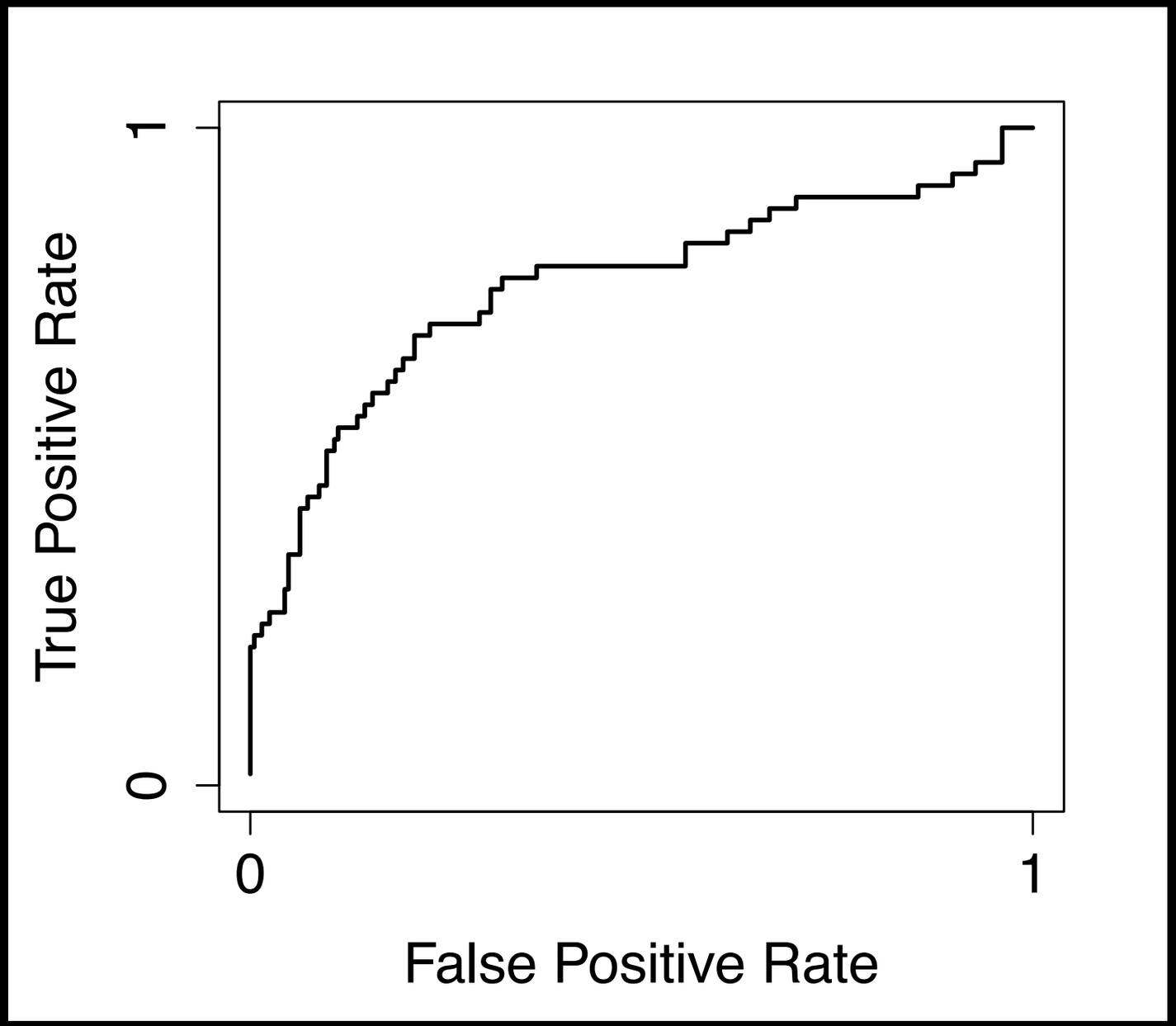
2. Use these predictions in an intelligent way
   to recommend a set of recommend to test next

We use Naive Bayes as our machine learning technique

- Naive Bayes is a statistical model often used for text classification (e.g., spam filters).
  - It is called “naive” because it makes an independence assumption.
  - Although it is naive, it often works really well.

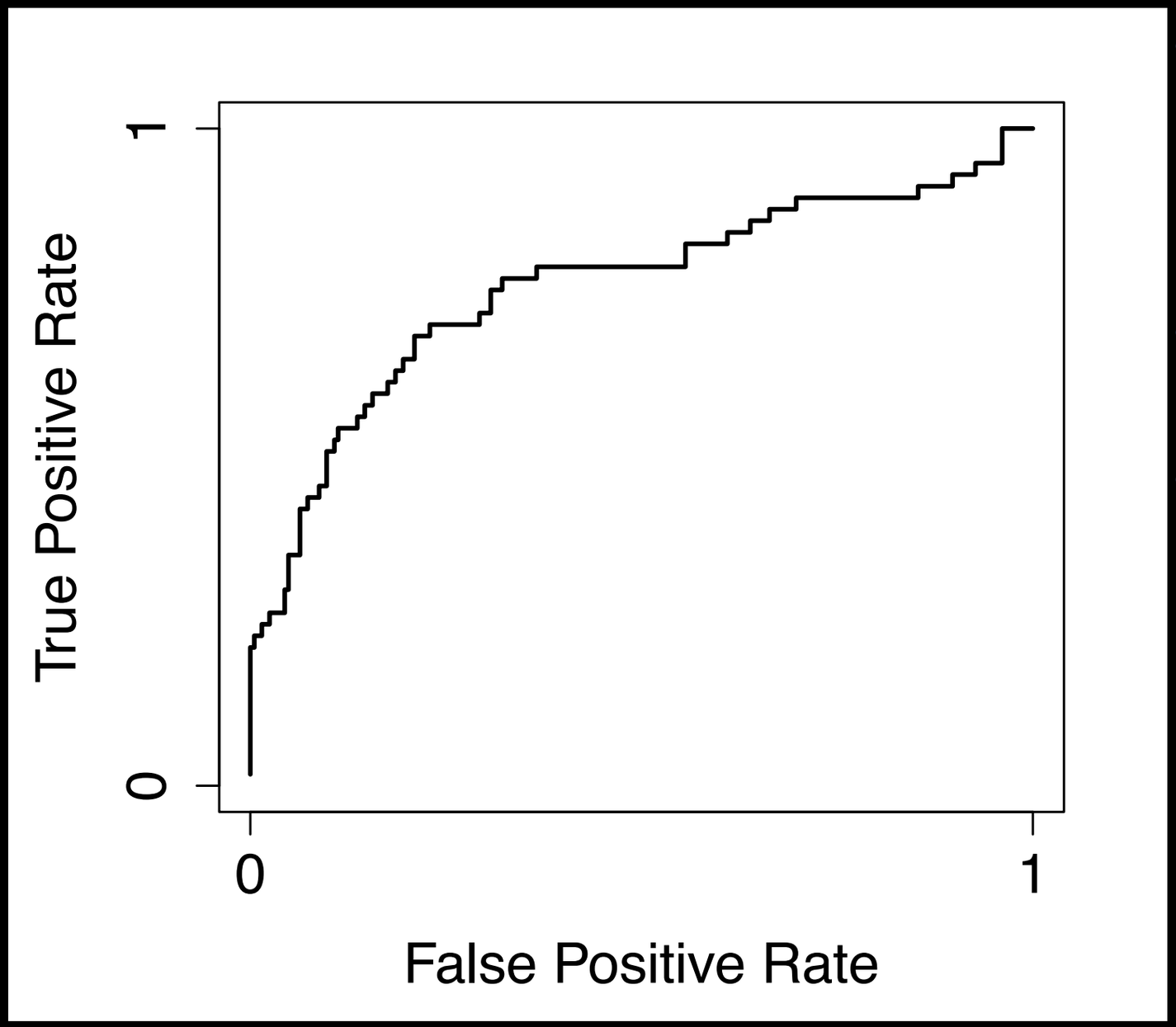
- We apply a variant of Naive Bayes to our problem, which is customized to include amino acids’ **location** within the peptide.

Naive Bayes is ok, but far from perfect



- This graph used training data from ~300 peptides (most are misses.)
- True positive rate = % of hits predicted to be hits.
- False positive rate = % of misses predicted to be hits.
- Rates were estimated via leave-one-out cross-validation.

Given imperfect predictions, what should we test next?



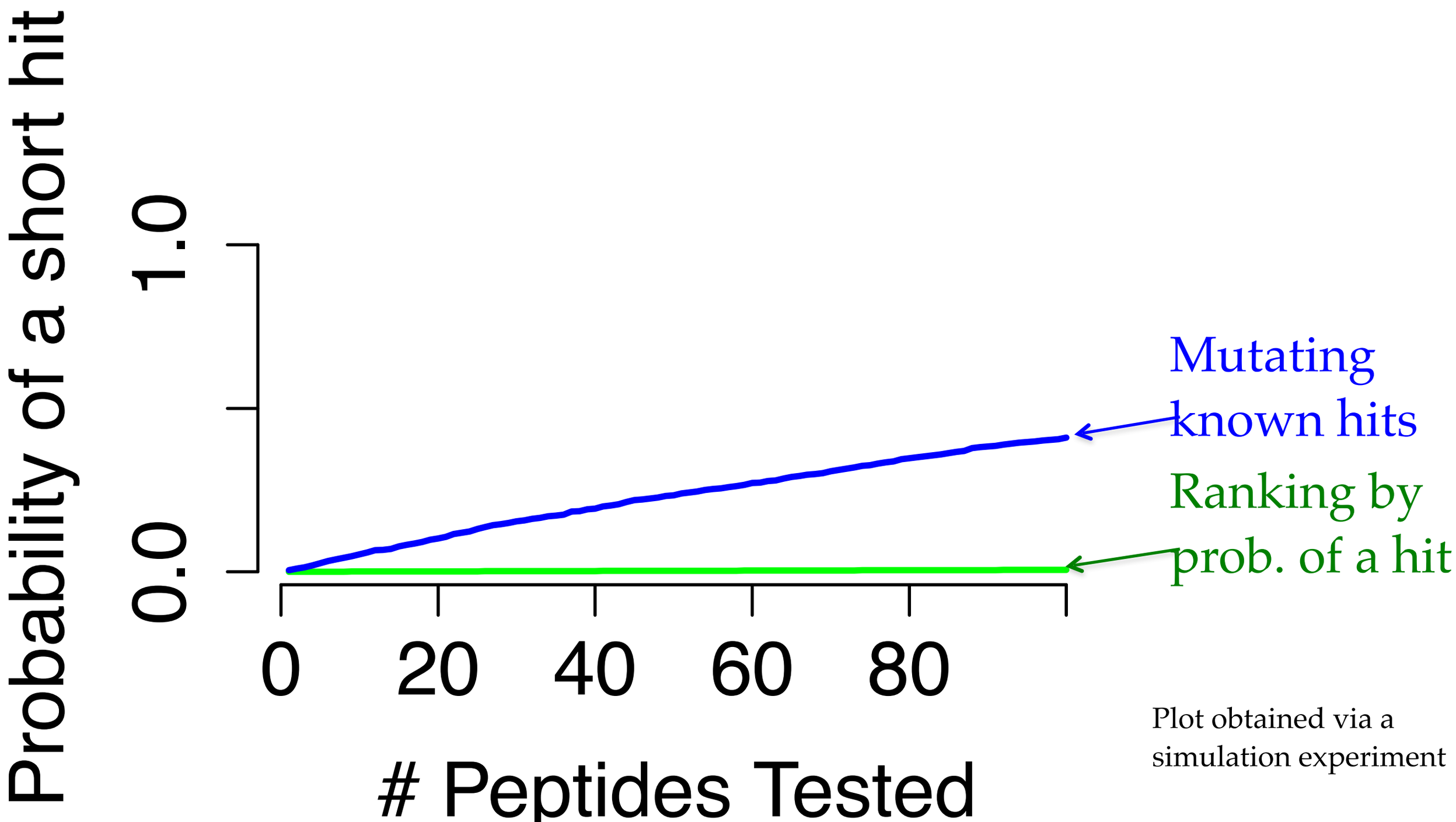
- If predictions were perfect, we could just test the shortest peptide predicted to be a hit.
- Our predictions are not perfect.
- How should we decide what to test next?

Ranking by probability of a hit does not work well

- One simple strategy is:
  - Select those peptides with length < 12.
  - Rank them by predicted probability of a hit
  - Test the top 300.

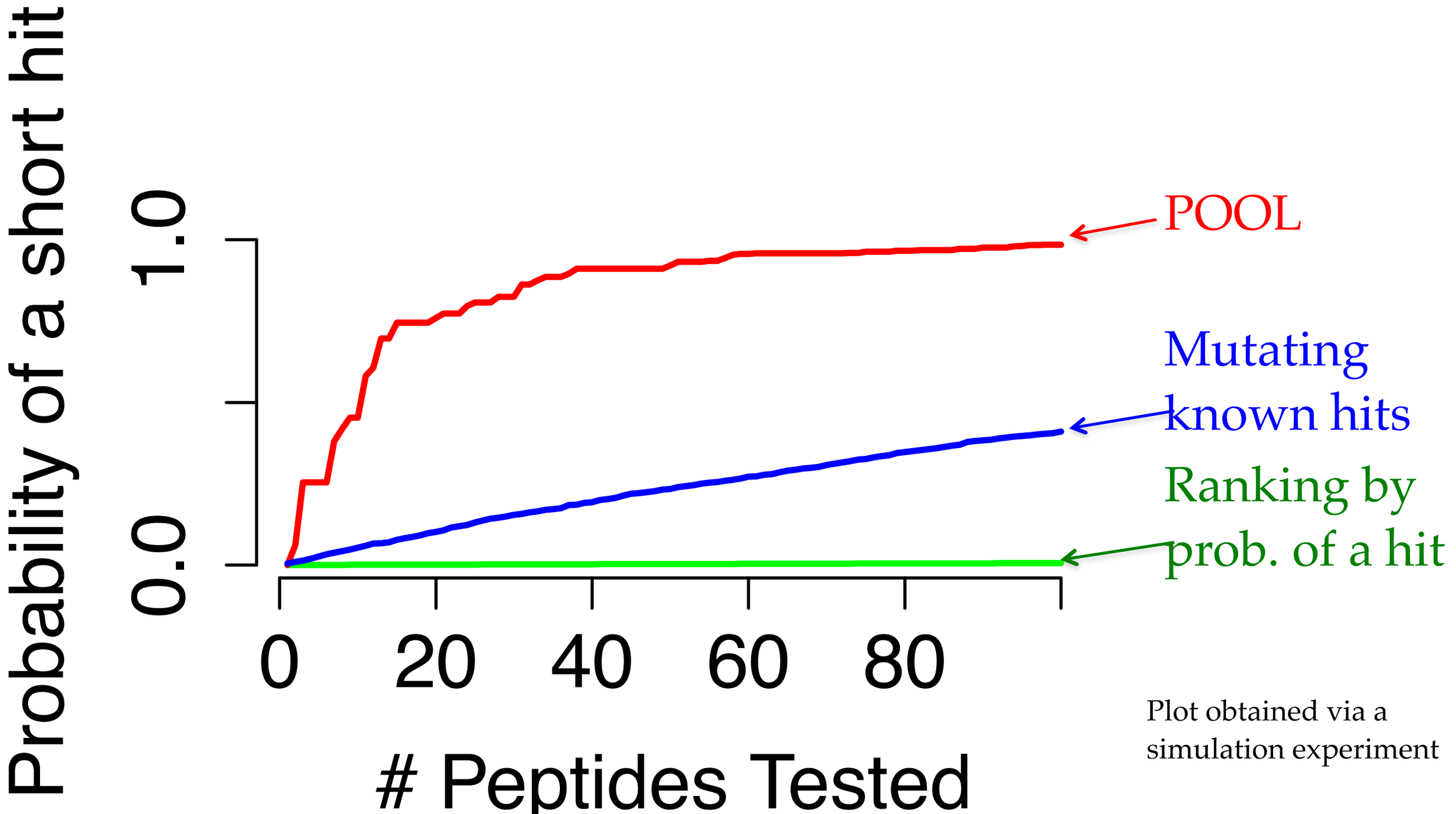
- The tested peptides are very similar. If the first tested peptide is not a hit, the other ones probably aren't either.

Ranking by probability of a hit does not work well



Plot obtained via a simulation experiment

POOL 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:

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

- 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 choose k.
  If b=14 and k=500, this is 10^{19} choose 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:

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

- Add e to S and repeat, until S has k=500 peptides.

The greedy algorithm performs within 63% of optimal

Let P*(S) = P(at least one short hit in S).

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

Proposition: Let 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

$$\frac{\text{OPT} - \text{GREEDY}}{\text{OPT}} \leq 1 - 1/e$$

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 the single peptide to add that maximizes the probability of reaching our goal:

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

- Is equivalent to:

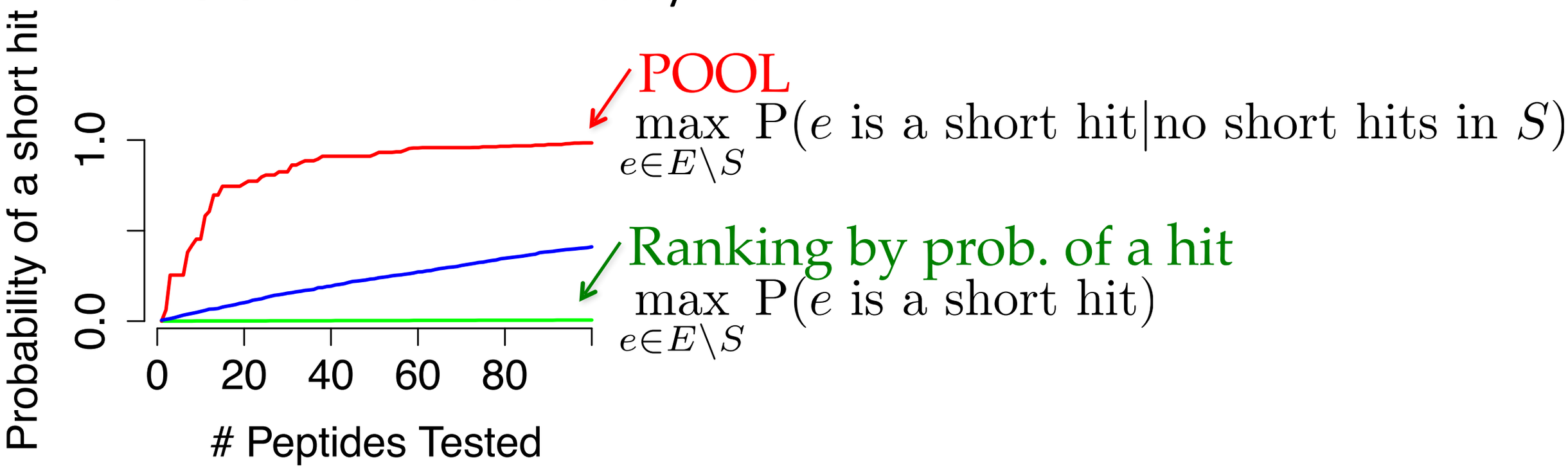
$$\max_{e \in E \setminus S} \mathrm{P}(e \text{ is a short hit} | \text{no short hits in } S)$$

- Compare this to the “rank by prob. hit” approach

$$\max_{e \in E \setminus S} \mathrm{P}(e \text{ is a short hit})$$

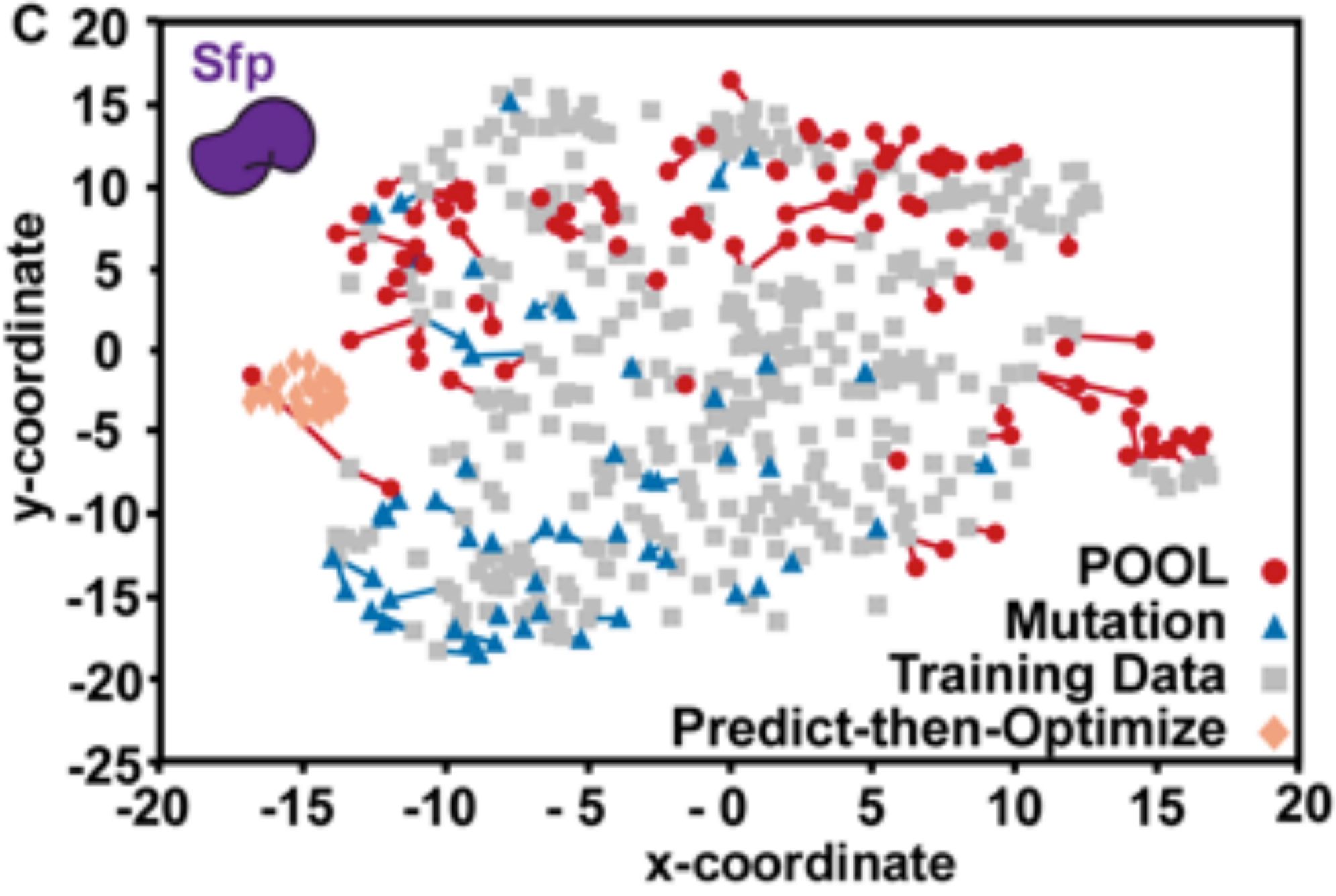
POOL works better because its peptides are more diverse

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



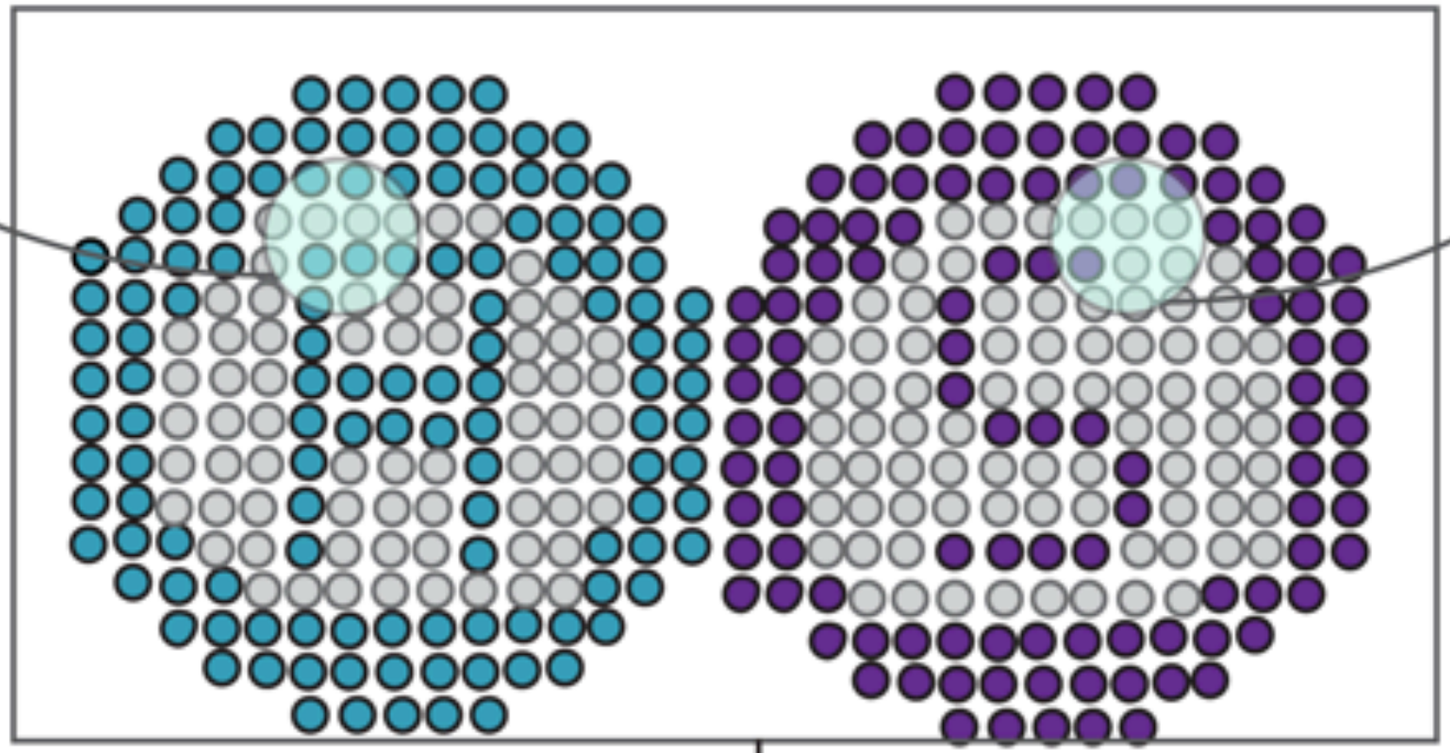
- Its recommendations are more **diverse**.

POOL's recommendations are more diverse



AcpS-specific hit
(w/ control)

Sfp-specific hit
(w/ control)

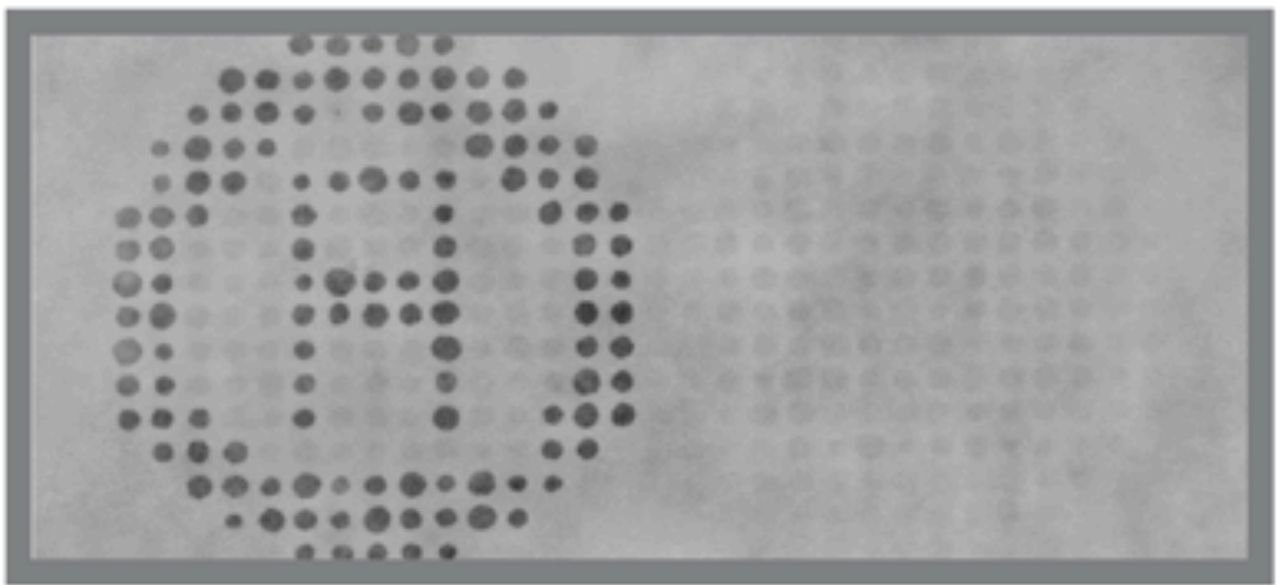


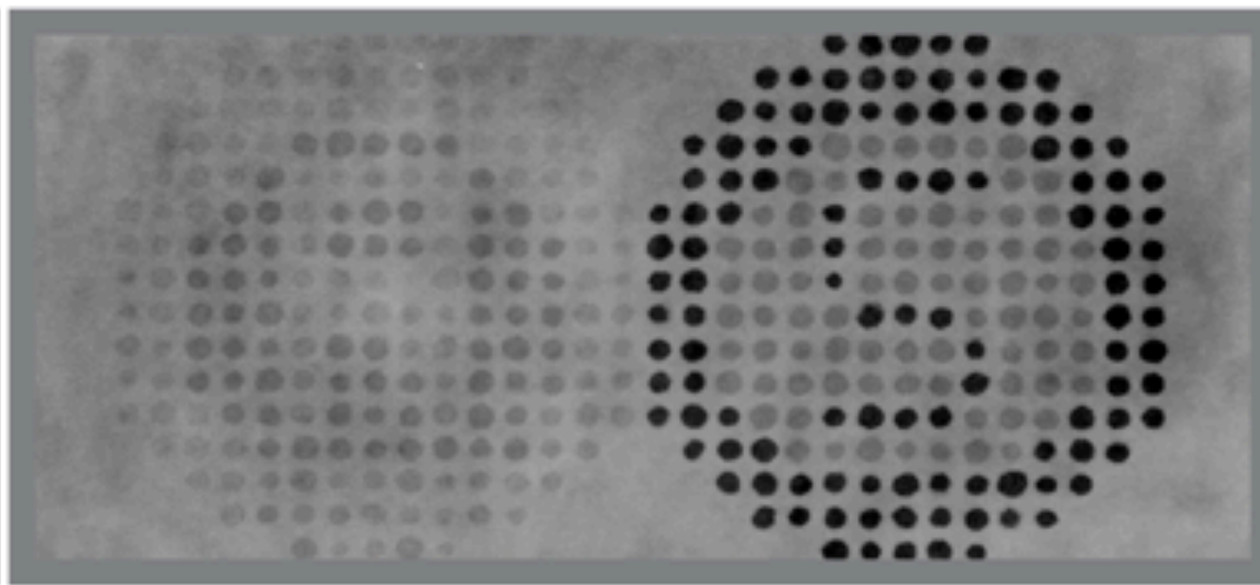
orthogonal labeling

λ_{ex} = 532 nm
λ_{ex} = 580 nm

AcpS

Sfp





Conclusion

- We have developed an optimal learning method for finding minimal peptide substrates.
- This method has found hits shorter than the shortest previously known.
- This approach can be applied to other kinds of problems.
- This approach aims to:
  - Reduce the experimental effort required to reach a goal.
  - Increase the chance of achieving a goal within a given experimental budget.

Thank you!

Appendix

We use Naive Bayes

- We assume that reality is characterized by a pair of latent matrices, called $\theta^{(\mathrm{hit})}$ and $\theta^{(\mathrm{miss})}$, where columns of each matrix correspond to different positions within the peptide, and rows correspond to different types of amino acids.

- These latent matrices are unknown, but can be estimated from data.

- We further suppose that, for a peptide x,

$$P(y(x) = 1 | x, \theta^{\mathrm{hit}}, \theta^{\mathrm{miss}}) = \frac{P(\mathrm{hit}) \prod_i \theta^{(\mathrm{hit})}_{i,x_i}}{P(\mathrm{hit}) \prod_i \theta^{(\mathrm{hit})}_{i,x_i} + P(\mathrm{miss}) \prod_i \theta^{(\mathrm{miss})}_{i,x_i}}$$

- Here, x is a peptide, x_i is the type of the amino acid at position i, y(x) indicates whether x is a hit (1) or not (0), and P(hit) and P(miss) are prior estimates of the fraction of hits and misses in the population.

We use Bayesian Naive Bayes

- We put independent Dirichlet prior distributions on each column of the latent matrices $\theta^{(\text{hit})}$ and $\theta^{(\text{miss})}$.

- Our choices for the parameters of this prior are based on a biological understanding of the problem, discussions with our collaborators, and cross validation.

- Given training data $x^1,\ldots,x^n$, $y(x^1),\ldots,y(x^n)$, the posterior on the θ's is also Dirichlet, and independent across i and j.

- To estimate the posterior probability of a hit, we can sample the θ's from the posterior, or calculate a single MAP estimate. The MAP estimate ignores uncertainty, but can be computed analytically.

Using VOI to optimize P(≥1 short hit) has a shortcoming

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

- Thus, the experiments that maximize P(≥1 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:

$$\mathrm{EI}(S) = E[(b - f^*(S))^+]$$

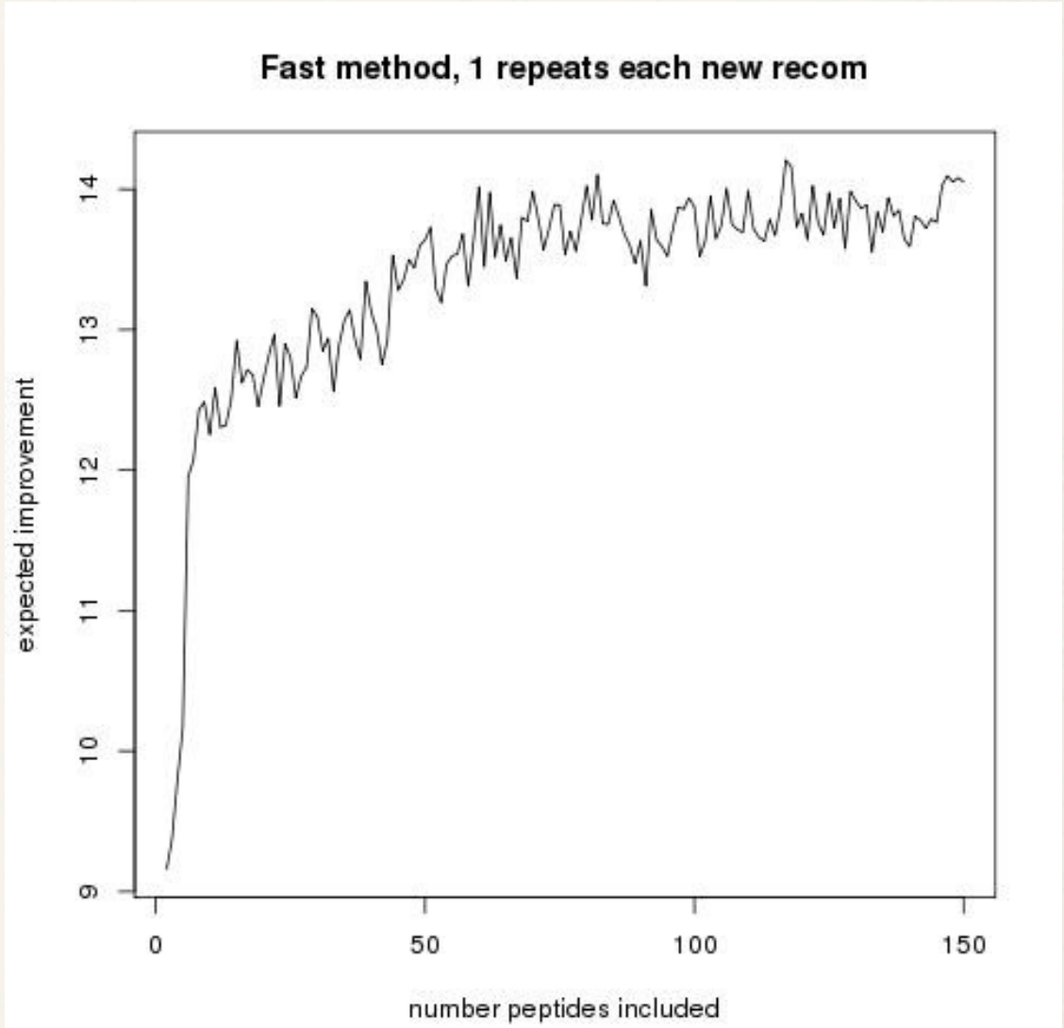
- An S that maximizes EI(S) could contain peptides shorter than b-1.

Efficiently optimizing expected improvement is ongoing work

- Solving $\max_{S \subseteq E : |S| \leq k} \mathrm{EI}(S)$ 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.

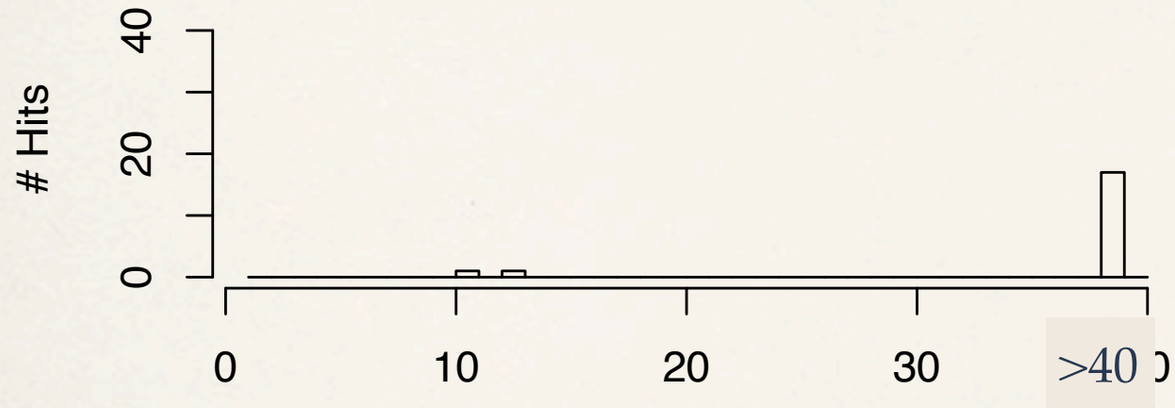
We are greedily optimizing P(≥1 short hit) with one tweak to make real recommendations

- 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(≥1 short hit).

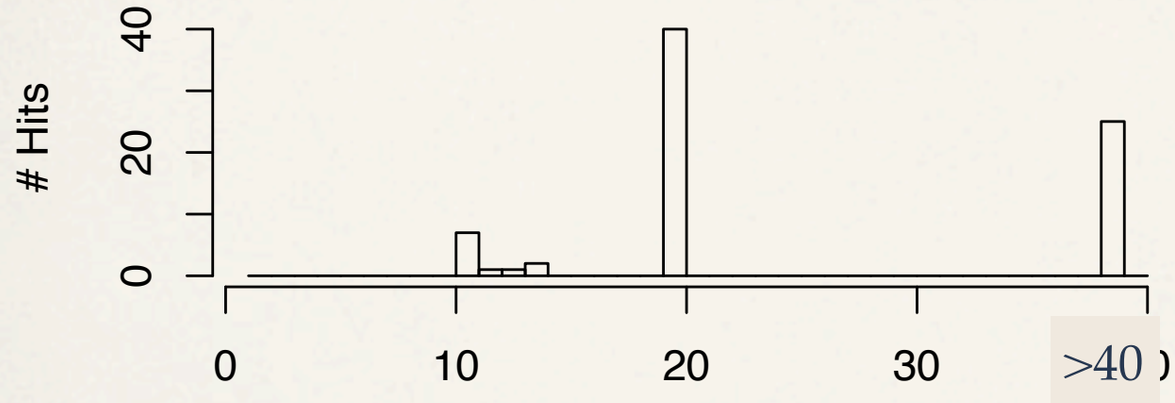


Expected improvement as a function of |S|, estimated via Monte Carlo.

We have found novel short peptides using this method



Training Set
Length of shortest hit: 11



After 1 round of POOL
Length of shortest hit: 11

After 2 rounds of POOL
Length of shortest hit: 10