Accelerating Scientific Discovery through Interpretable Machine Learning & Intelligent Experimentation

Peter Frazier

Associate Professor, Operations Research & Information Engineering, Cornell

Staff Data Scientist, Uber

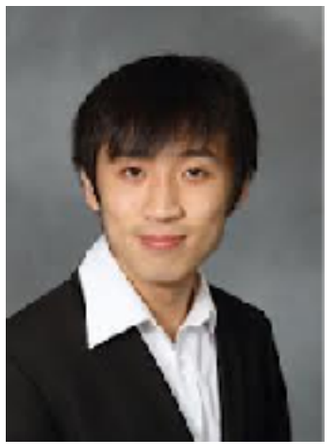
Lori Tallorin
(UCSD)

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)

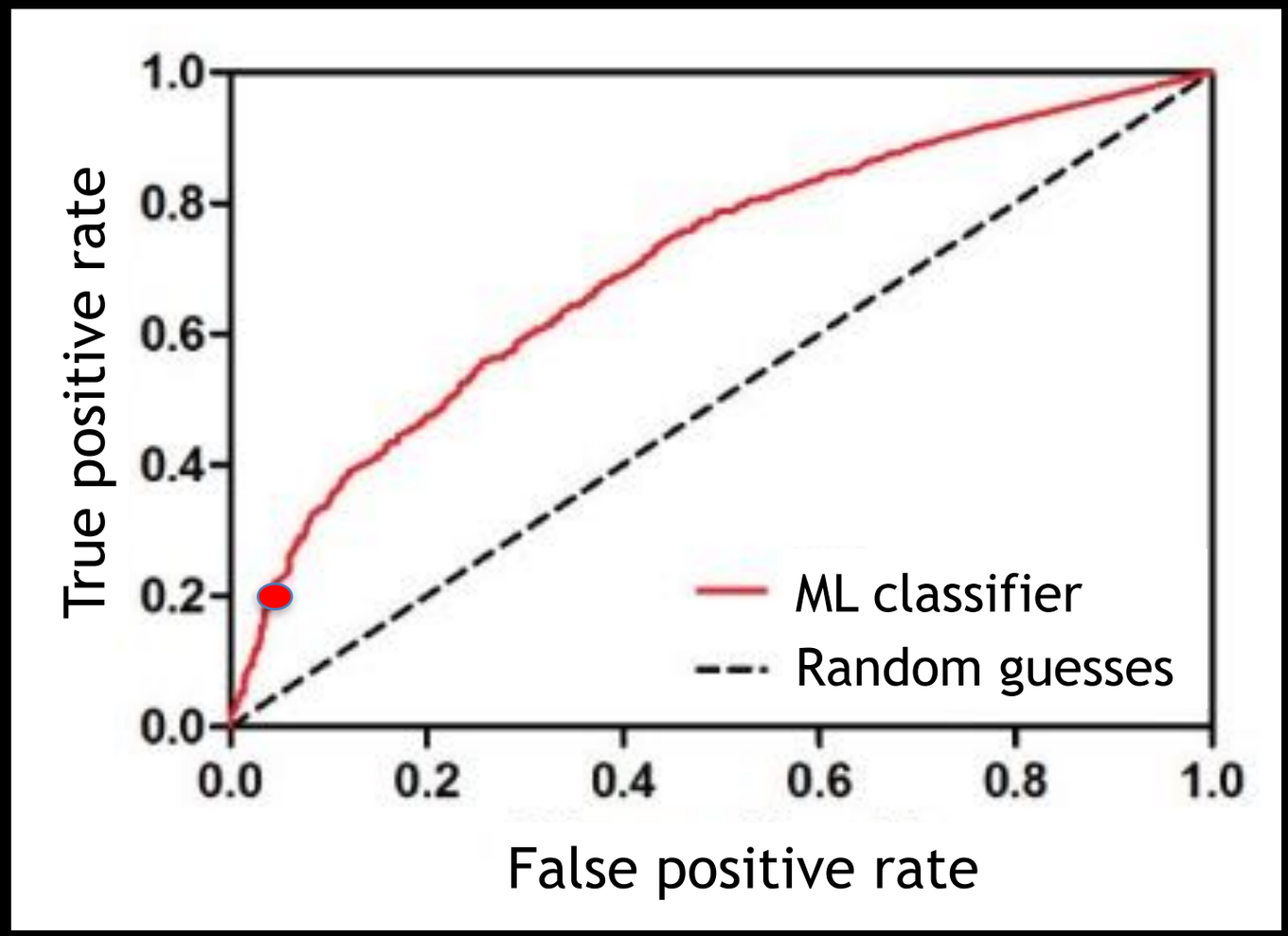
"Discovering *de novo* peptide substrates for enzymes using machine learning." *Nature Communications*, 2018

Here's one way to use ML for peptide design

1. Collect experimental training data
2. Train a supervised learning model
3. *In silico*, rank peptides by predicted activity, and take the top 10
4. Evaluate activity for these top 10 peptides

Challenge:

Machine learning makes errors,
especially with small amounts of training data



At the red dot, we label:

5% of non-active peptides as active

20% of active peptides as active

If 1 in 10^5 peptides are active,
we'll have to test >10^4 predicted-active peptides to find the first active one.

Opportunities to improve ML's effectiveness

- Get more training data
- Build prediction methods that perform well with less data
- Make good decisions about which experiments to perform (Bayesian optimization; Active Learning)

Opportunities to improve ML's effectiveness

- Get more training data

Interpretable models

- Build prediction methods that perform well with less data

- Make good decisions about which experiments to perform (Bayesian optimization; Active Learning)

Opportunities to improve ML's effectiveness

- Get more training data
- Build prediction methods that perform well with less data
- Make good decisions about which experiments to perform (Bayesian optimization; Active Learning)

Opportunities to improve ML's effectiveness

- Get more training data
- Build prediction methods that perform well with less data
- Make good decisions about which experiments to perform (Bayesian optimization; Active Learning)

Interpretable Bayesian optimization

Example:

Recommender Systems

Step 1: Use ML 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?

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

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

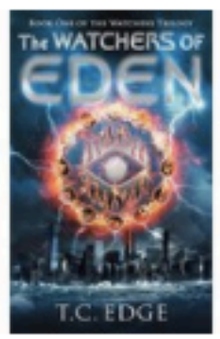
Paperback
$11.99 ~~$12.99~~ Prime
Get it by **Saturday, Dec 10**

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 like 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

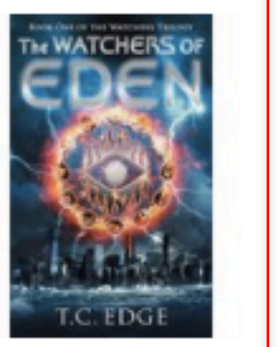
Kindle Edition
$0.00
Whispersync for Voice-ready

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

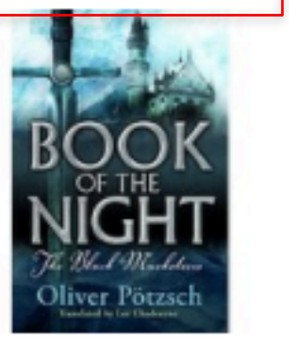
Paperback
$13.99 Prime
Get it by **Saturday, Dec 10**

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

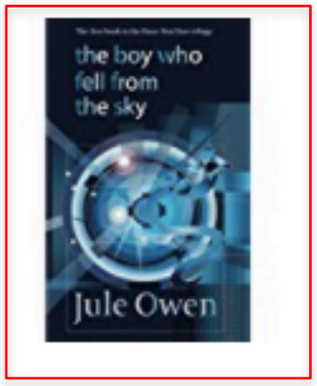
Kindle Edition
$0.00 kindleunlimited
Read this and over 1 million books with Kindle Unlimited.

$3.99 to buy
Whispersync for Voice-ready

Step 2: Take ML'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

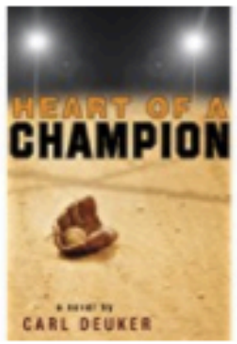
Kindle Edition
$0.00
Whispersync for Voice-ready

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

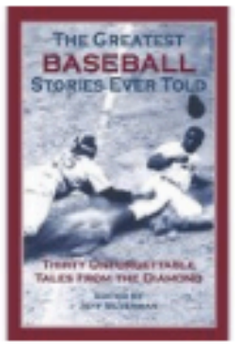
Kindle Edition
$9.99
Auto-delivered wirelessly

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

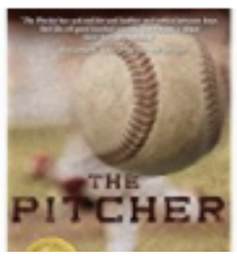
Paperback
$13.12 ~~$16.95~~ Prime
Get it by **Saturday, Dec 10**

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 ML'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

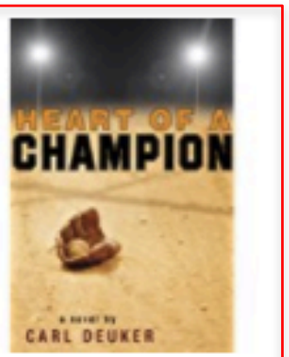
Kindle Edition
$0.00
Whispersync for Voice-ready

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

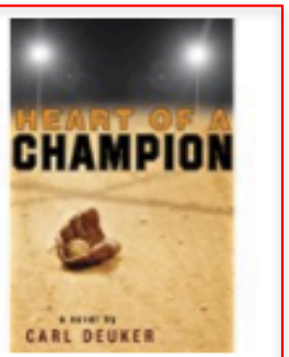
Kindle Edition
$0.00
Whispersync for Voice-ready

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

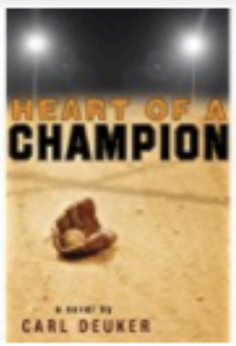
Kindle Edition
$0.00
Whispersync for Voice-ready

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%



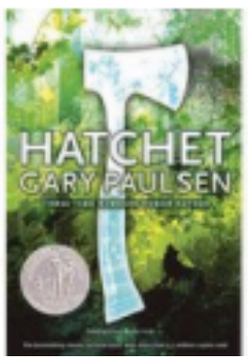
Heart of a Champion May 30, 2009
by Carl Deuker

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 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

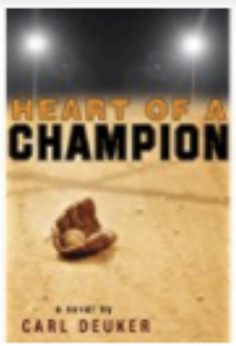
Kindle Edition
$0.00
Whispersync for Voice-ready

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%



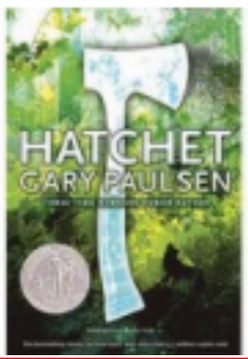
Heart of a Champion May 30, 2009
by Carl Deuker

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 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**

Providing a diverse selection increases the chance he'll like at least one from 48.7% to 1-.6*.5*.8=76%

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

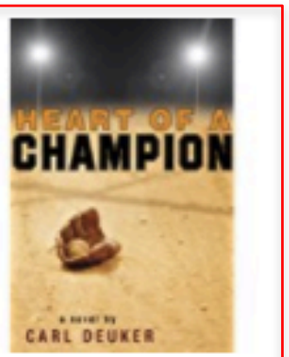
Kindle Edition
$0.00
Whispersync for Voice-ready

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%



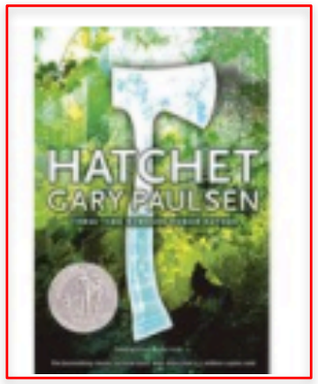
Heart of a Champion May 30, 2009
by Carl Deuker

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 the first two

20%



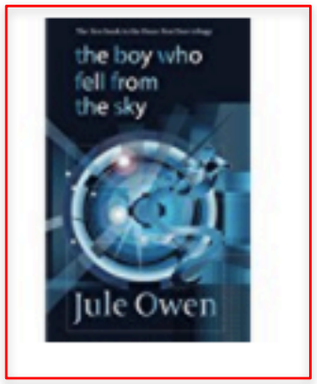
Hatchet (Brian's Saga Book 1) Aug 25, 2009
by Gary Paulsen

Kindle Edition
$6.99
Whispersync for Voice-ready

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

We do 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.

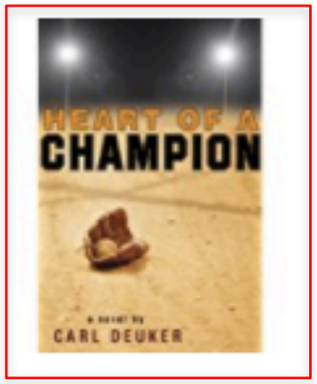


The Boy Who Fell from the Sky (The House Next Door Book 1) Sep 22, 2015
by Jule Owen

Kindle Edition
$0.00
Whispersync for Voice-ready

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

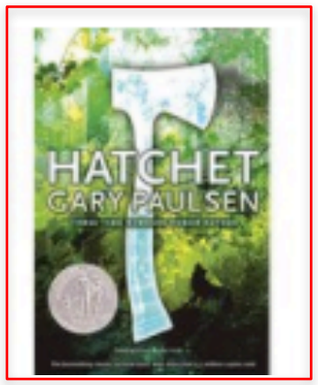
More Buying Choices



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 Bayesian optimization

Bayesian optimization optimizes time-consuming-to-evaluate functions.

Bayesian optimization iterates 2 steps:

1. Build a Bayesian supervised learning model of the objective
2. Suggest experiments to run based on an acquisition function

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

We are using Bayesian optimization to develop orthogonal protein labels

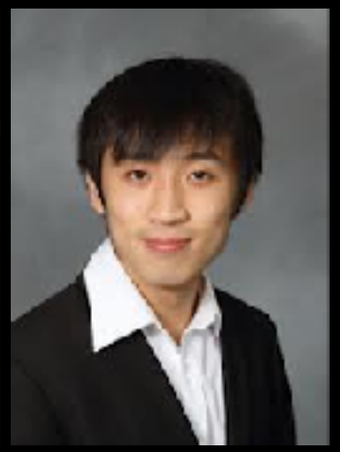
Lori Tallorin (UCSD)

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)

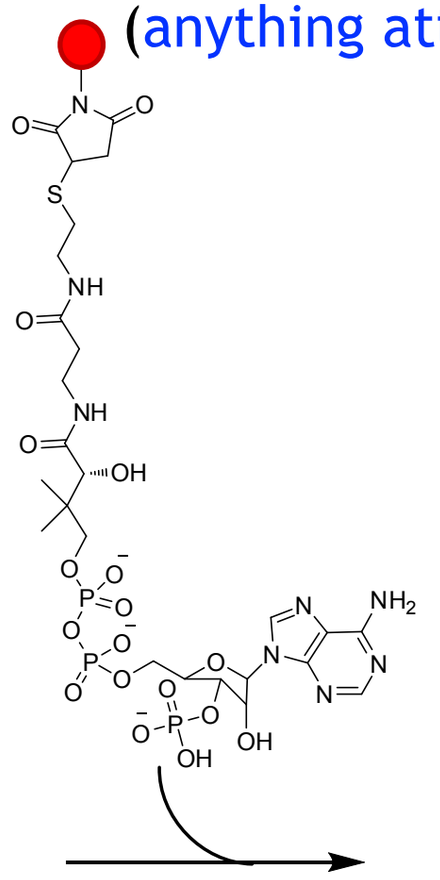
Mike Burkart (UCSD)

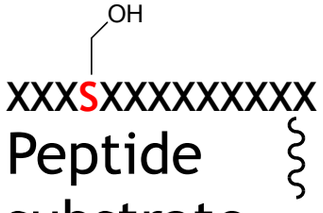
Our goal is to build a way to stick things to proteins

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

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

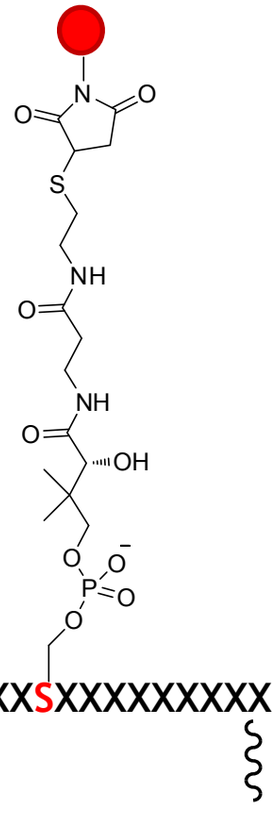
Phosphopantetheine
(pin)





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

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

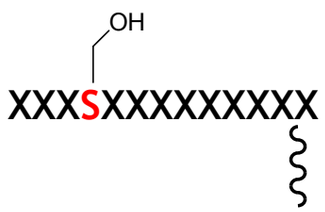
  (1) an Sfp-specific peptide substrate labeled by Sfp but not by AcpS

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

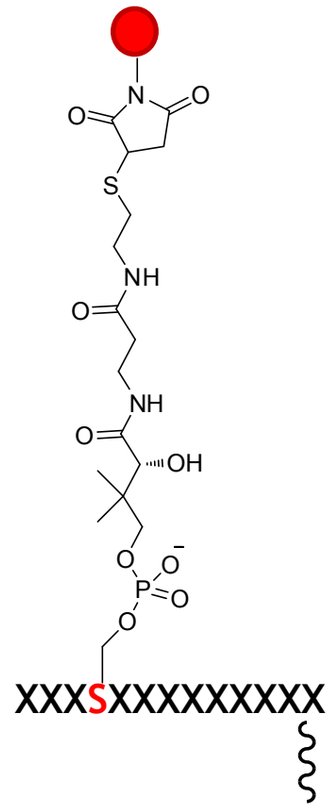
arbitrary label, e.g., red dye
(anything attached to pin)

Phosphopantetheine
(pin)

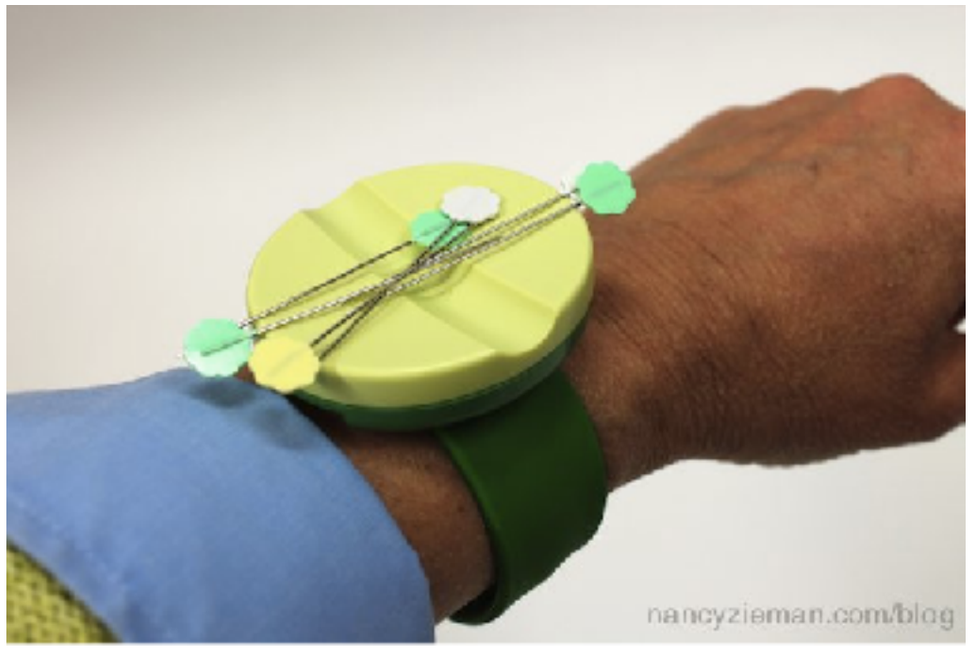
Peptide
substrate
(pincushion),
can be embedded in large protein
(person wearing the pincushion)



PPTase enzyme
(person sticking pin into pincushion)



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



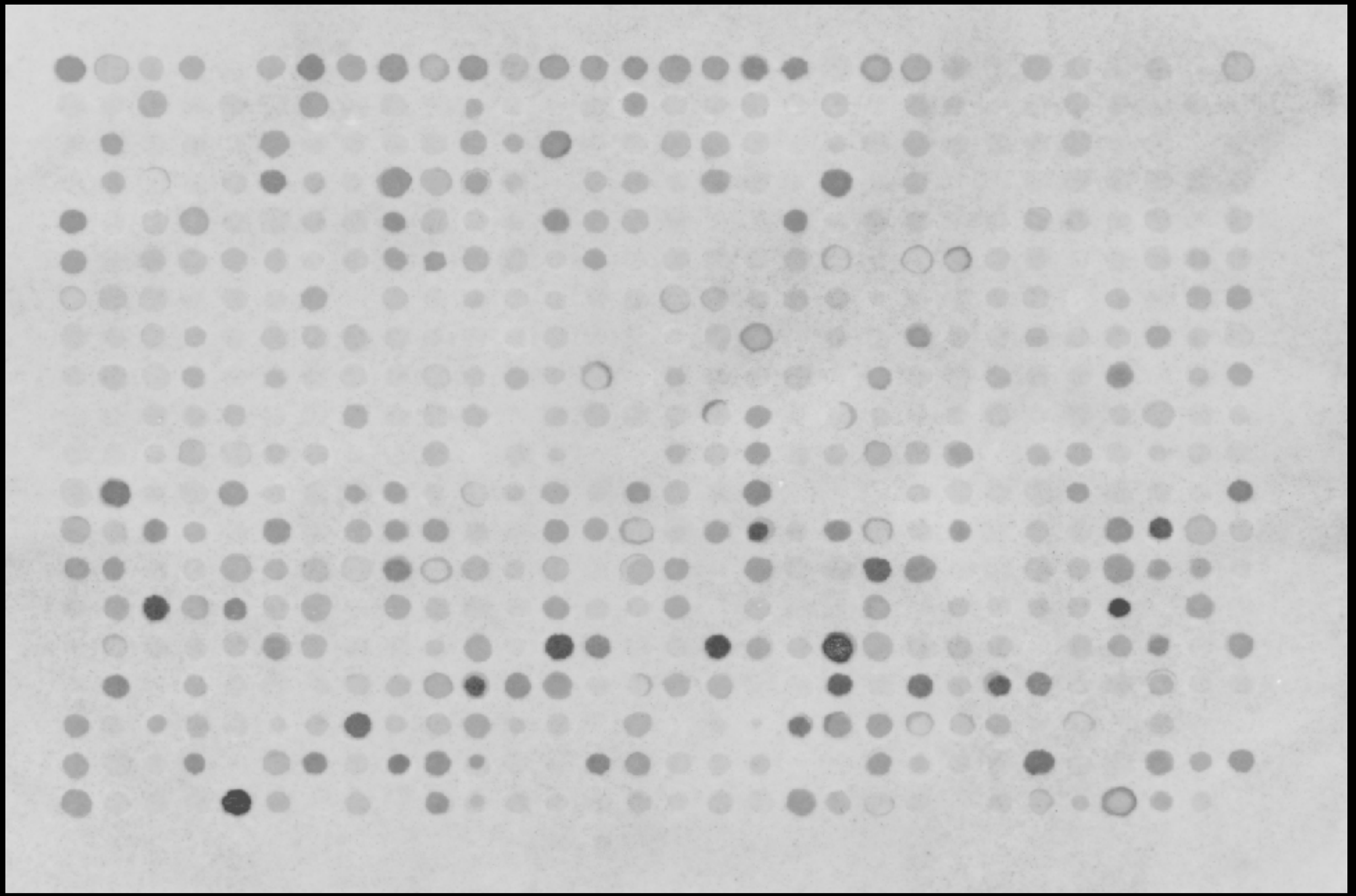
To make our orthogonal labeling system useful, we need the substrates to be short

- If a peptide is a substrate for Sfp and not AcpS, we call it a “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.

- 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



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 (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: It's not deep
- Advantages:
  - Easy to explain to collaborators
  - Easy to understand & debug
  - Easy to customize a prior to our application
  - Strong prior => Robust to extremely small amounts of data
  - Good quantification of uncertainty
  - Computational scalability!

Naive Bayes assumes two latent matrices

$\theta^{(\text{hit})}$, P(amino acid | 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	.06	.32	.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	.08
FWY	.08	.08	.27	.27	.07	.07	.15	.06	.08	.04		.04	.05	.11	.06	.06	.07	.07	.07	.07	.24
HKR	.10	.08	.12	.07	.10	.15	.06	.06	.05	.04		.04	.05	.06	.06	.07	.07	.11	.07	.16	.08
AILMV	.21	.12	.10	.27	.20	.07	.46	.09	.61	.04		.61	.21	.14	.58	.19	.45	.33	.34	.29	.28
GP	.08	.08	.07	.07	.07	.07	.06	.54	.05	.04		.04	.05	.06	.06	.06	.07	.14	.09	.07	.08
ST	.19	.40	.14	.10	.07	.07	.06	.06	.05	.04		.15	.18	.46	.06	.17	.14	.14	.21	.07	.10
C	.08	.07	.07	.07	.07	.07	.06	.06	.05	.04		.04	.05	.06	.06	.06	.07	.07	.07	.07	.08

$\theta^{(\text{miss})}$, P(amino acid | 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	.06	.26	.07	.07	.12	.07	.08
NQ	.18	.08	.13	.07	.07	.16	.07	.06	.05	.08		.04	.19	.06	.06	.12	.07	.07	.07	.12	.08
FWY	.08	.08	.08	.27	.16	.16	.13	.06	.13	.04		.04	.05	.12	.06	.06	.07	.07	.07	.10	.18
HKR	.08	.13	.13	.07	.16	.07	.06	.06	.05	.08		.04	.12	.15	.06	.18	.07	.17	.07	.30	.08
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	.07	.10	.55	.17	.04		.04	.12	.06	.06	.07	.06	.07	.12	.07	.08
ST	.23	.13	.24	.07	.07	.07	.13	.06	.05	.12		.04	.05	.12	.09	.12	.23	.07	.17	.07	.08
C	.08	.08	.07	.07	.07	.07	.06	.06	.05	.04		.04	.05	.06	.06	.06	.07	.07	.07	.07	.08

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

We put independent Dirichlet priors on columns in these matrices

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

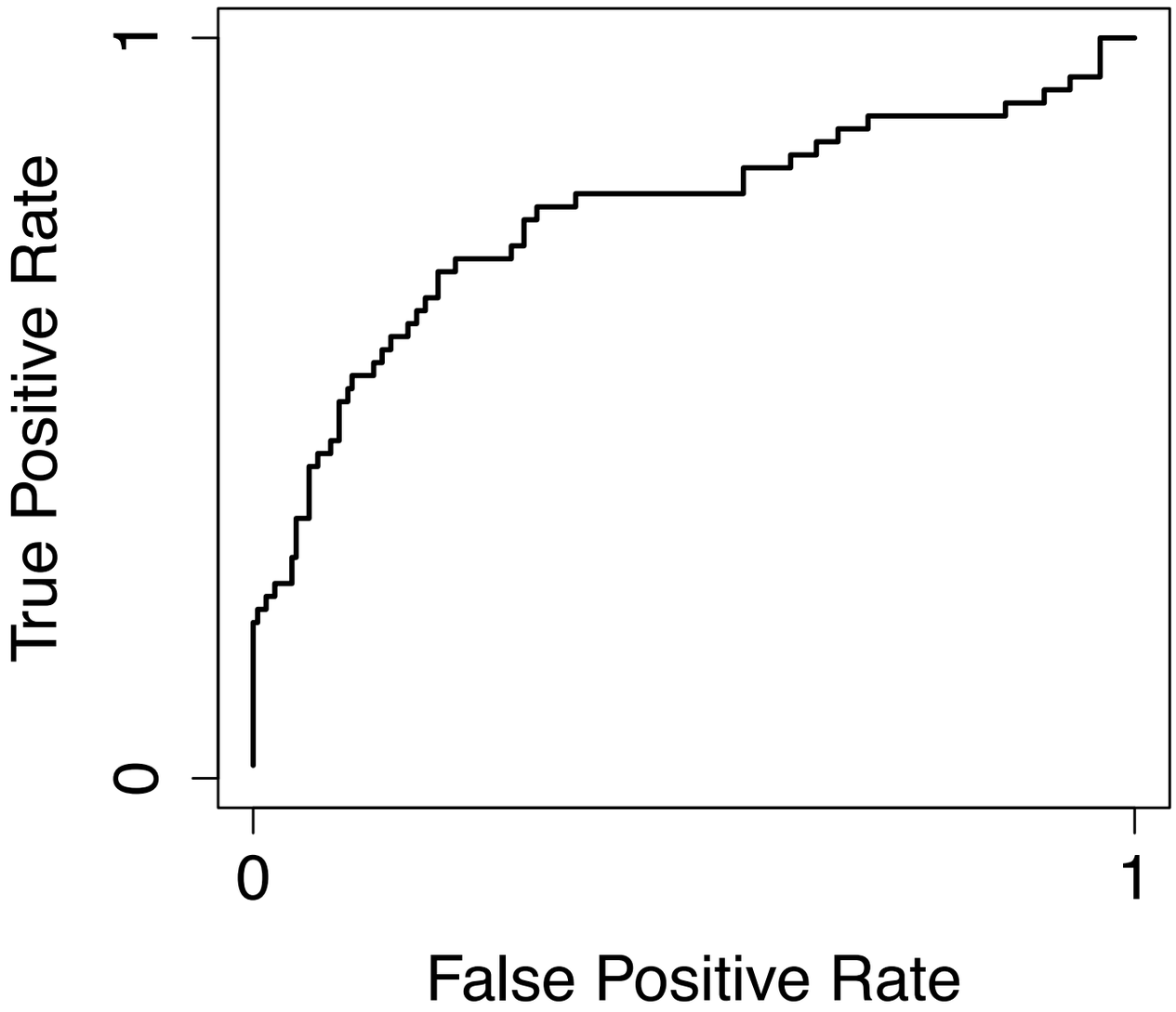
	Position relative to Serine																			
	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10
DE	.19	.08	.07	.07	.28	.45	.07	.09	.05	.72		.04	.37	.06	.06	.32	.07	.07	.07	.14	.08
NQ	.08	.10	.15	.07	.14	.07	.07	.06	.05	.04		.04	.05	.06	.06	.06	.07	.07	.07	.12	.08
FWY	.08	.08	.27	.27	.07	.07	.15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

$\theta^{(\text{miss})}$, P(amino acid | 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	.06	.26	.07	.07	.12	.07	.08
NQ	.18	.08	.13	.07	.07	.16	.07	.06	.05	.08		.04	.19	.06	.06	.12	.07	.07	.07	.12	.08
FWY	.08	.08	.08	.27	.16	.16	.13	.06	.13	.04		.04	.05	.12	.06	.06	.07	.07	.07	.10	.18
HKR	.08	.13	.13	.07	.16	.07	.06	.06	.05	.08		.04	.12	.15	.06	.18	.07	.17	.07	.30	.08
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	.07	.10	.55	.17	.04		.04	.12	.06	.06	.07	.06	.07	.12	.07	.08
ST	.23	.13	.24	.07	.07	.07	.13	.06	.05	.12		.04	.05	.12	.09	.12	.23	.07	.17	.07	.08
C	.08	.08	.07	.07	.07	.07	.06	.06	.05	.04		.04	.05	.06	.06	.06	.07	.07	.07	.07	.08

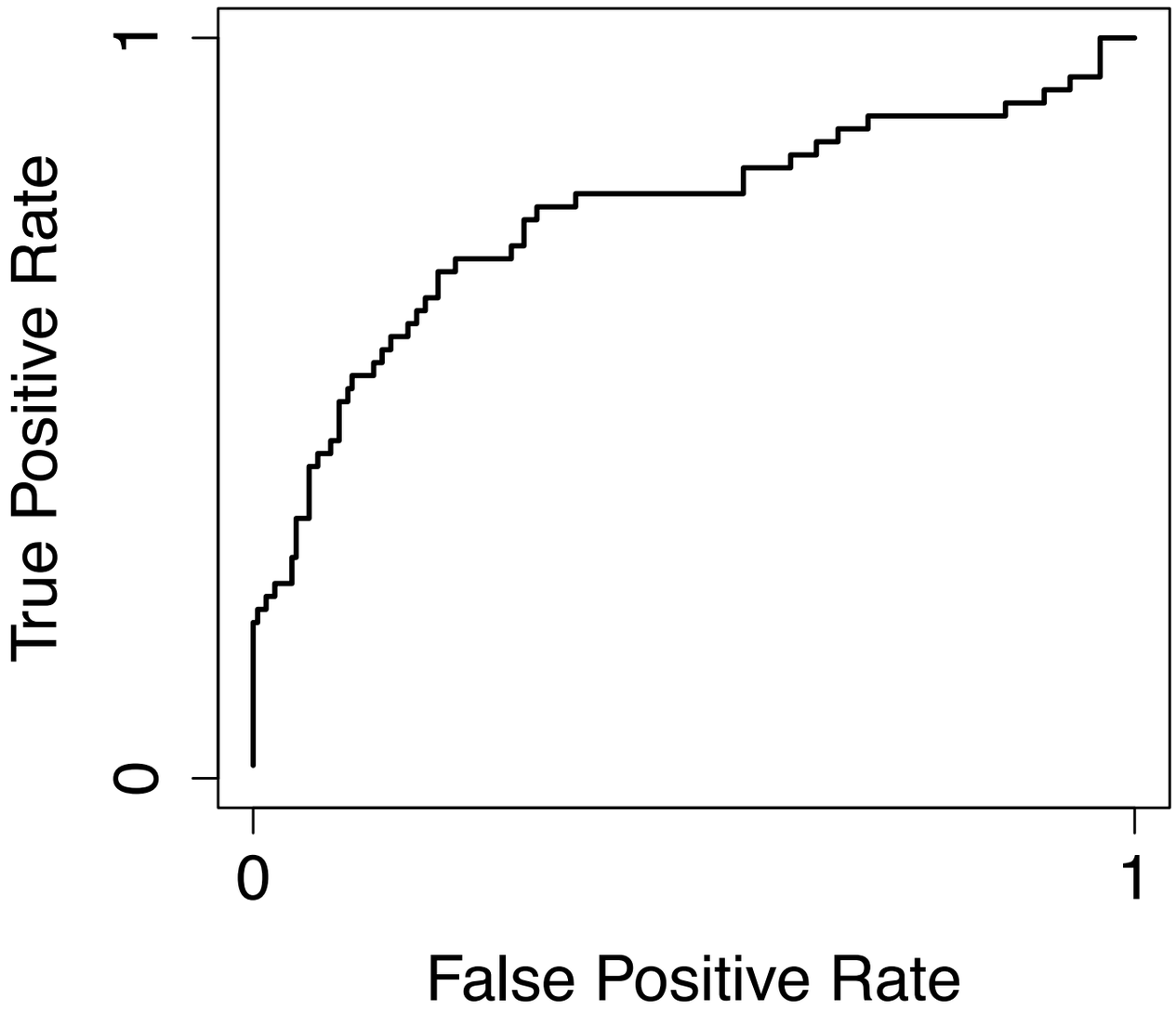
- Prior mean is proportional to the number of AA in the class
- The prior on columns far from the Serine is more concentrated close to the mean

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?



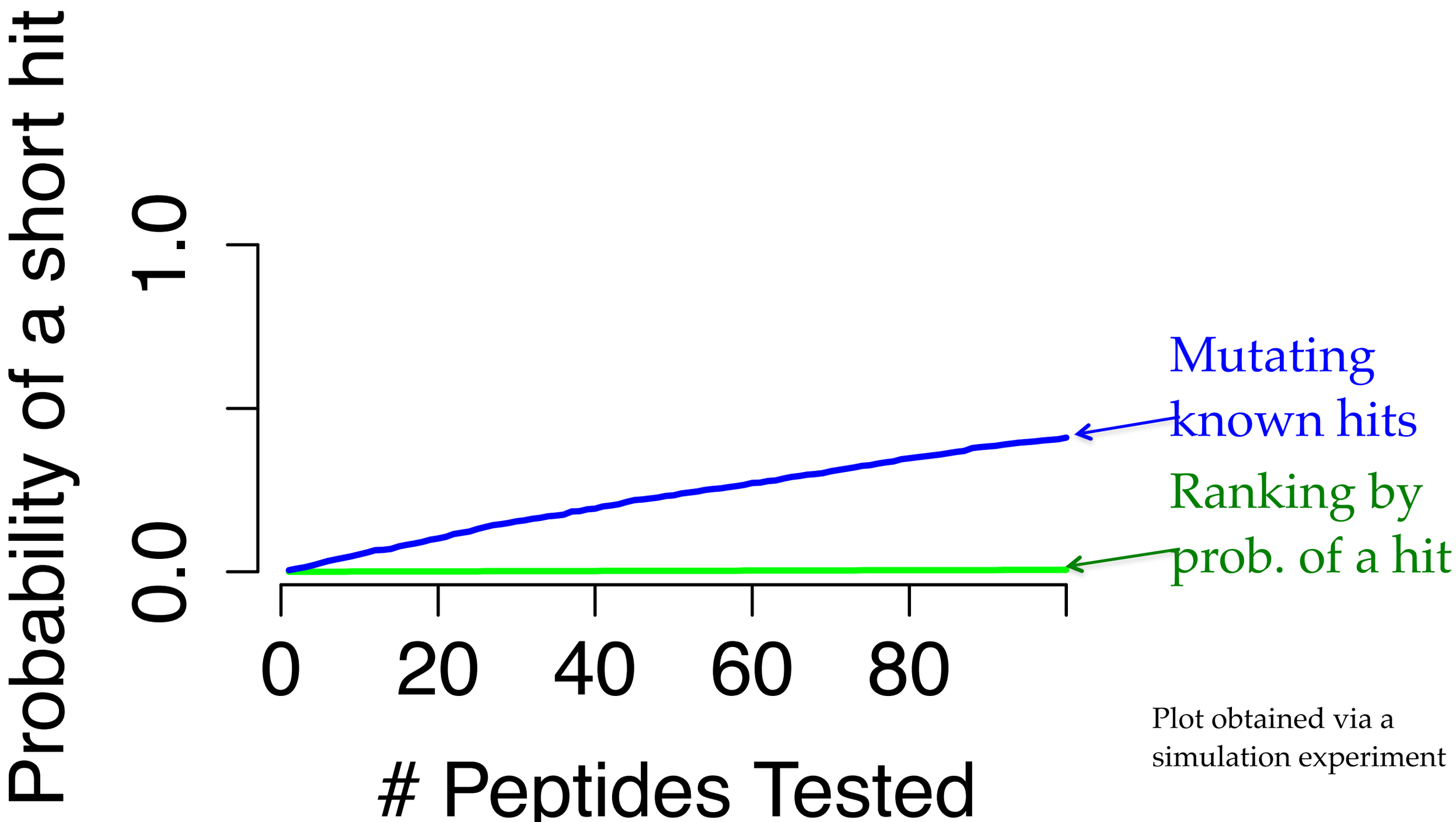
- If predictions were perfect, we could just test the shortest peptide predicted to be a hit.
- Our predictions are not perfect.
- How should we decide what to test next?

Ranking by probability of a hit does not work well

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

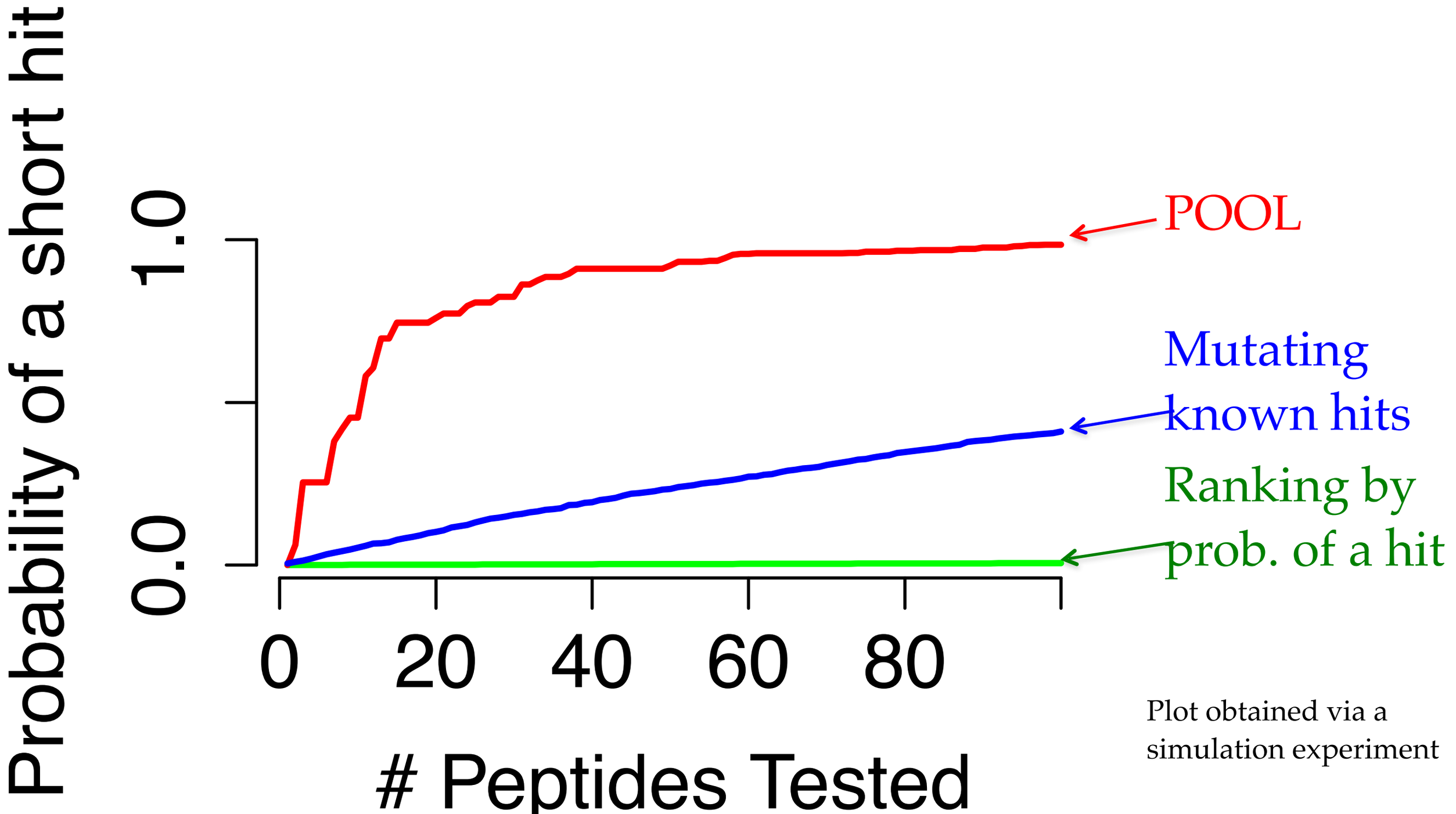
- The tested peptides are very similar. If the first tested peptide is not a hit, the other ones probably aren’t either.

Ranking by probability of a hit does not work well



Plot obtained via a simulation experiment

POOL 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.

* This isn't really our full goal, but it's pretty close.

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:

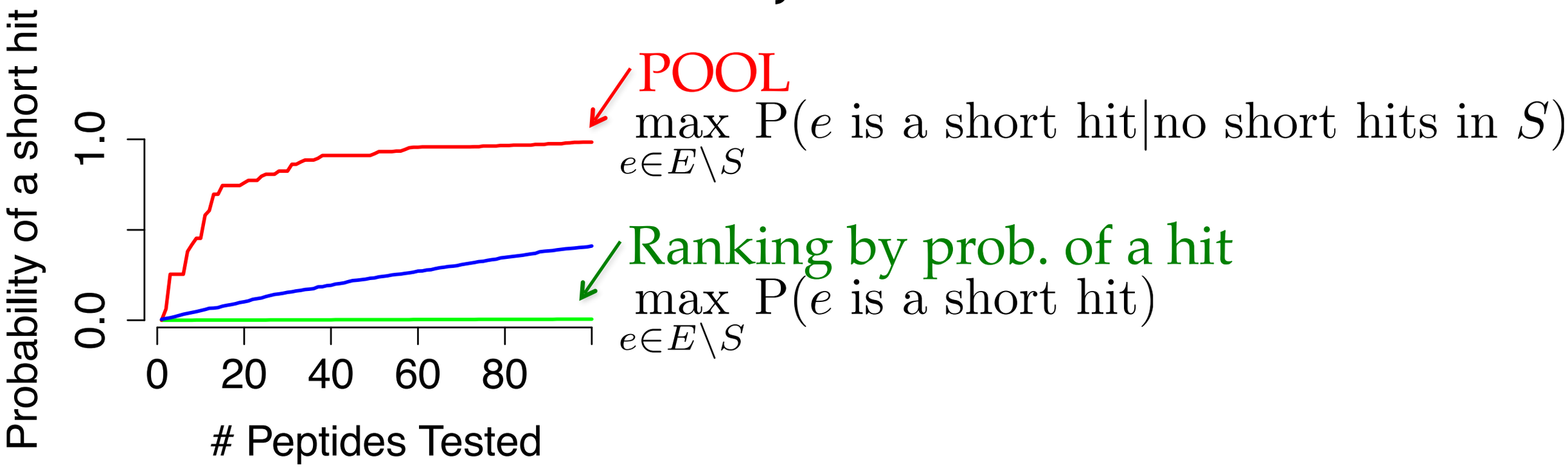
$$\max_{e \in E \setminus S} \mathrm{P}(e \text{ is a short hit} | \text{no short hits in } S)$$

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

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

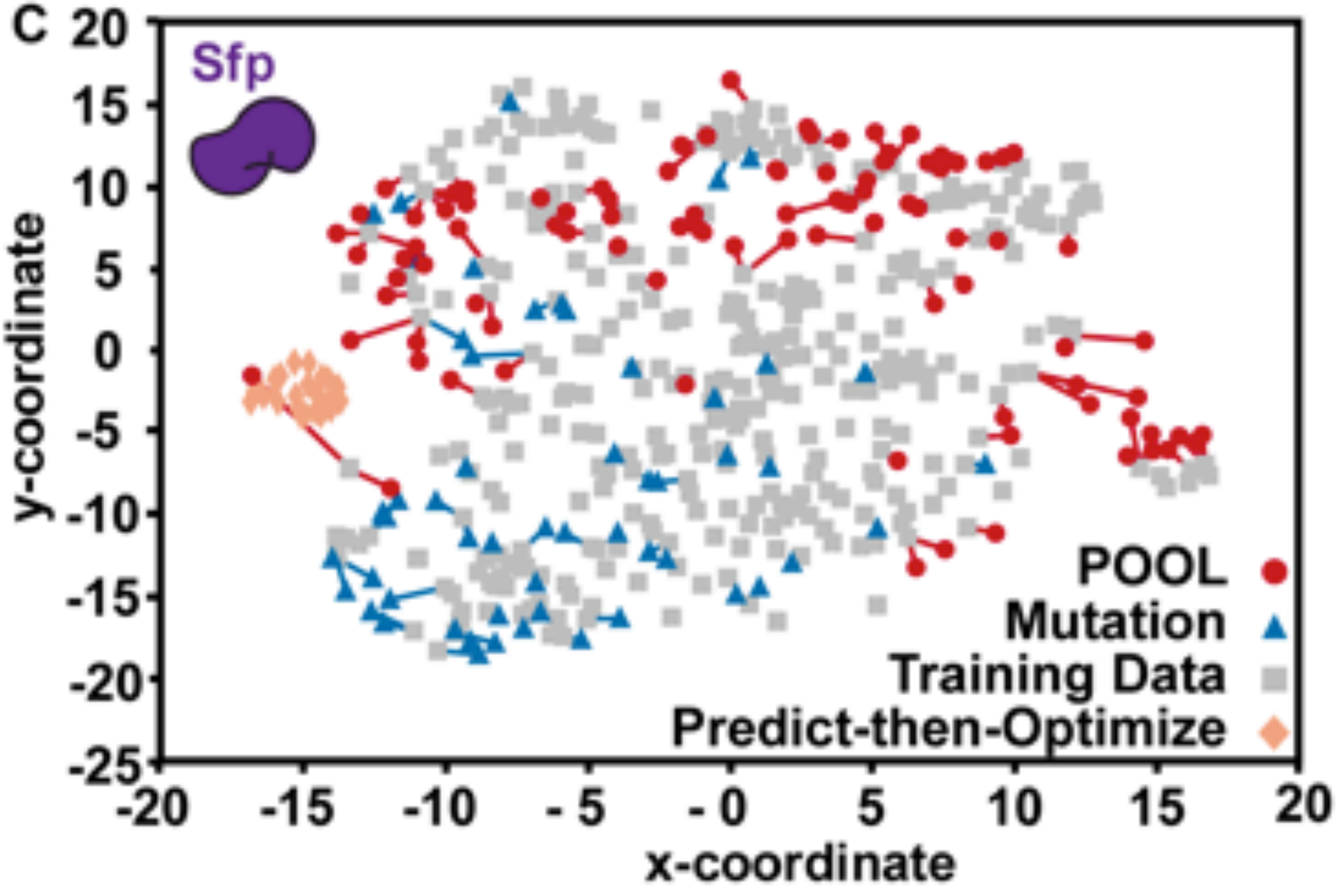
POOL works better because its peptides are more diverse

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

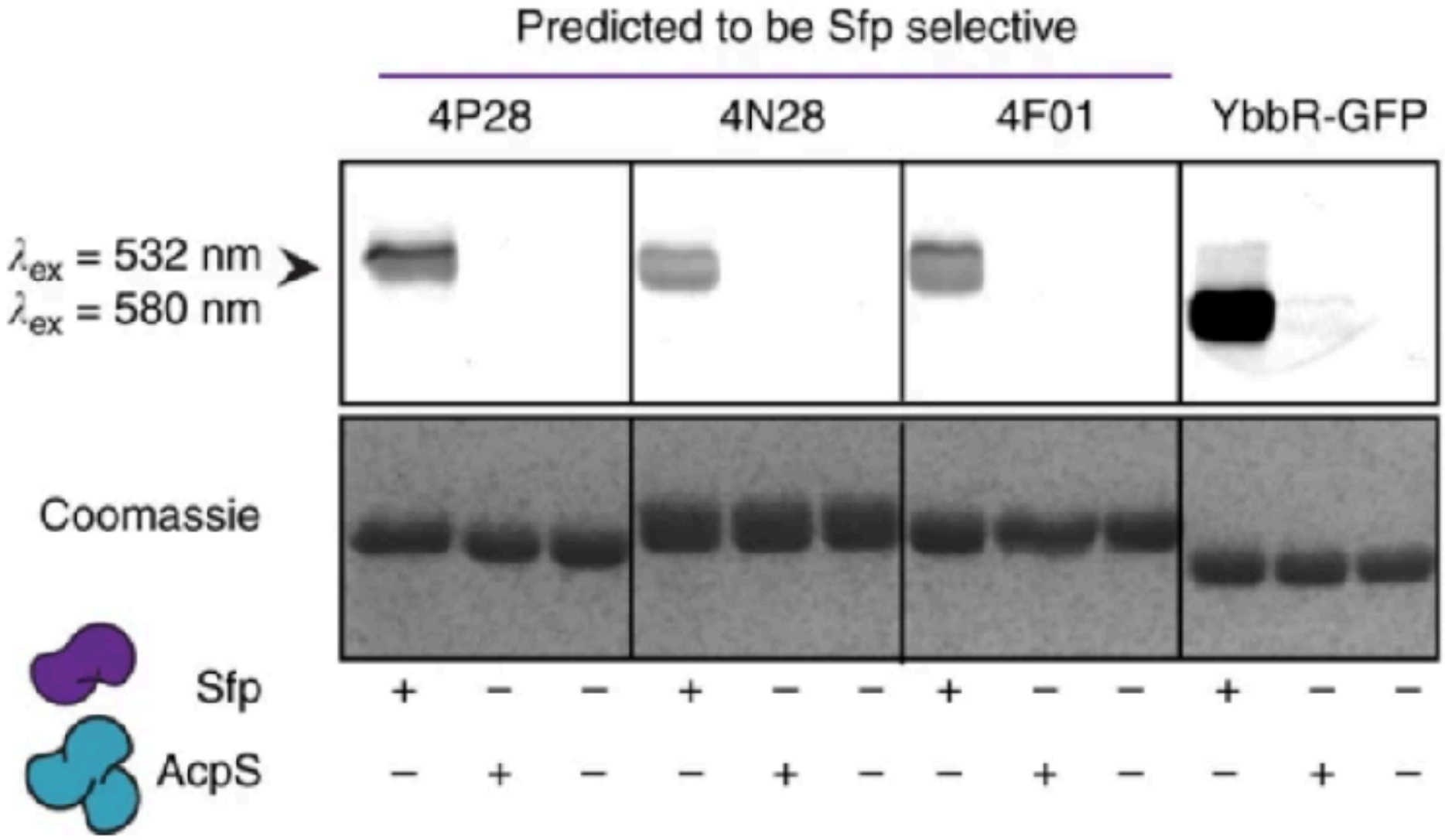


- Its recommendations are more **diverse.**

POOL's recommendations are more diverse



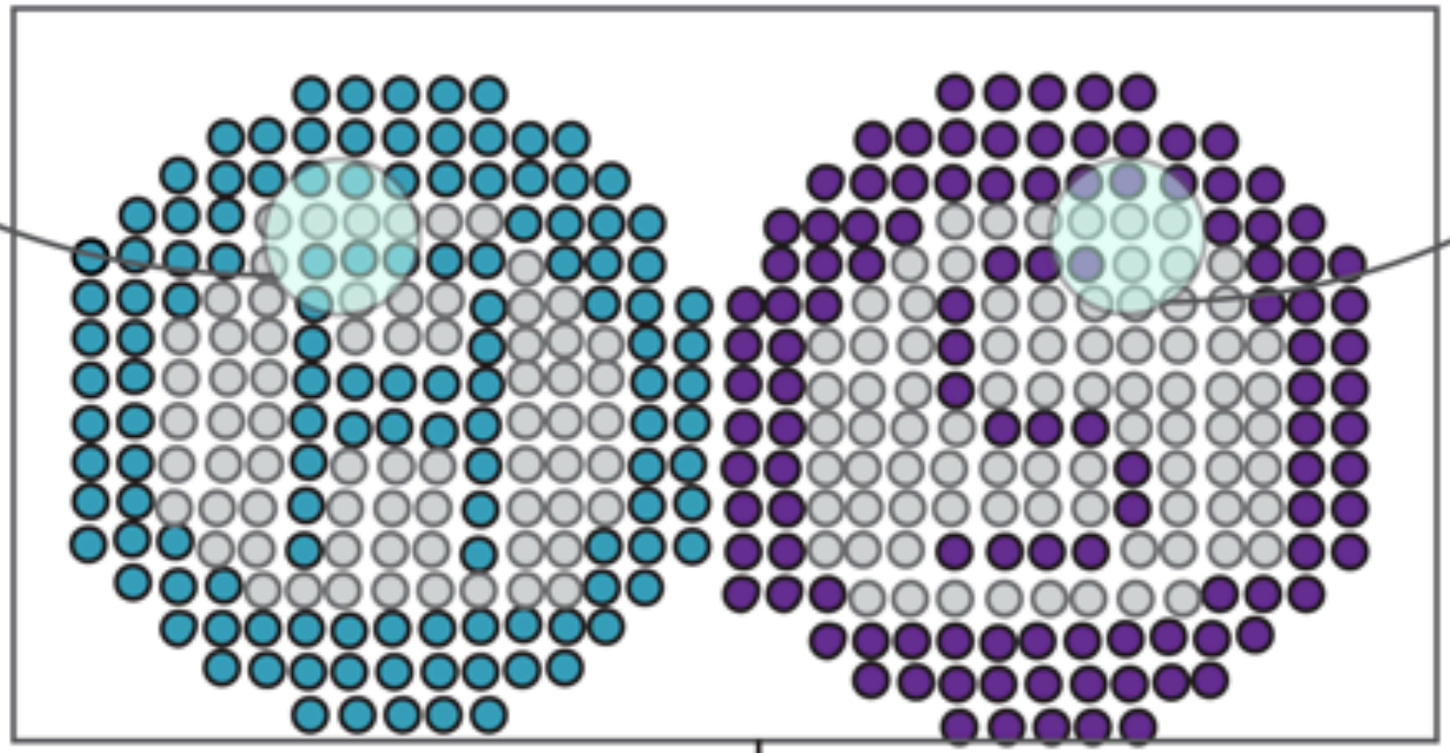
Sfp-type peptides can also be selectively labeled off-membrane, conjugated to GFP



We believe we were unable to label our AcpS-type GFP-peptides because of endogeneous AcpS in *E. coli* used to make them

AcpS-specific hit
(w/ control)

Sfp-specific hit
(w/ control)

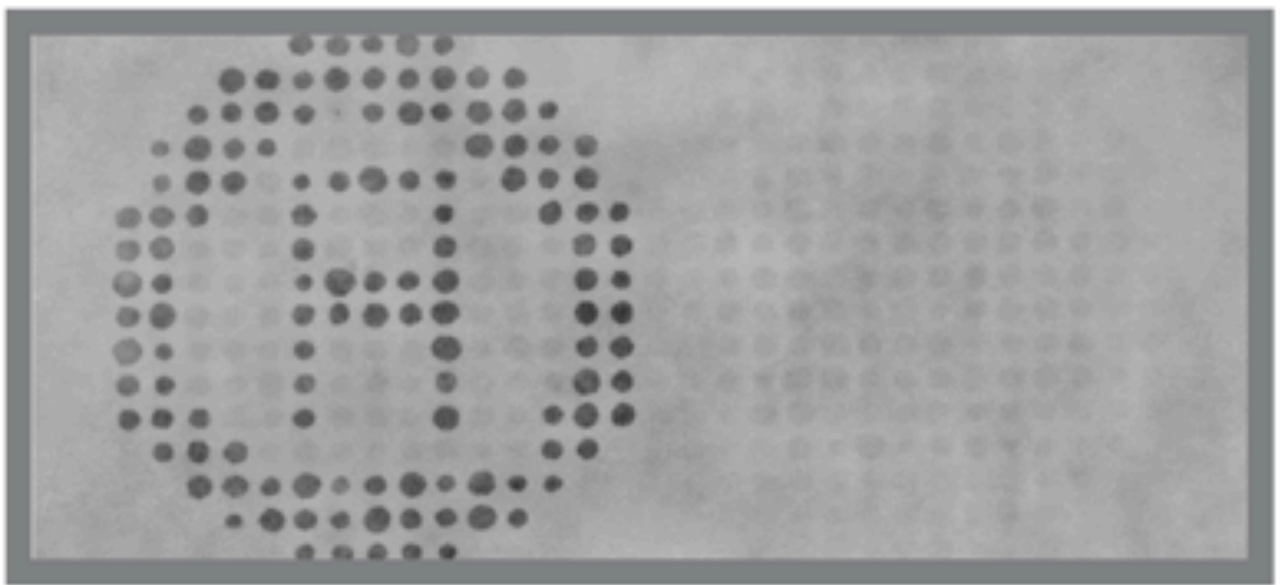


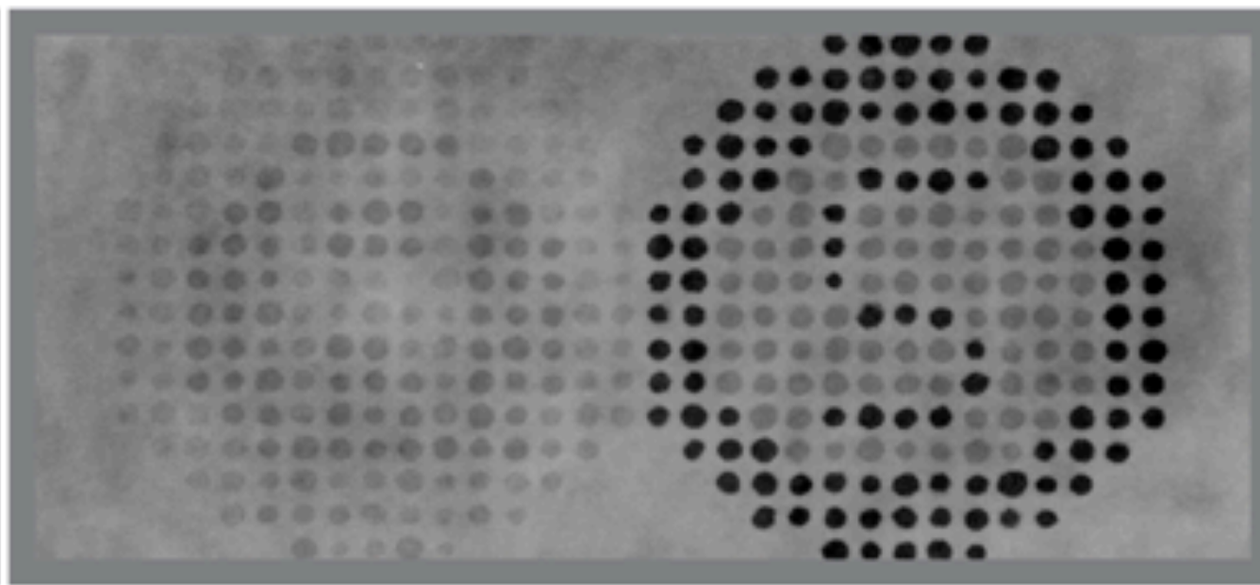
orthogonal labeling

λ_{ex} = 532 nm
λ_{ex} = 580 nm

AcpS

Sfp



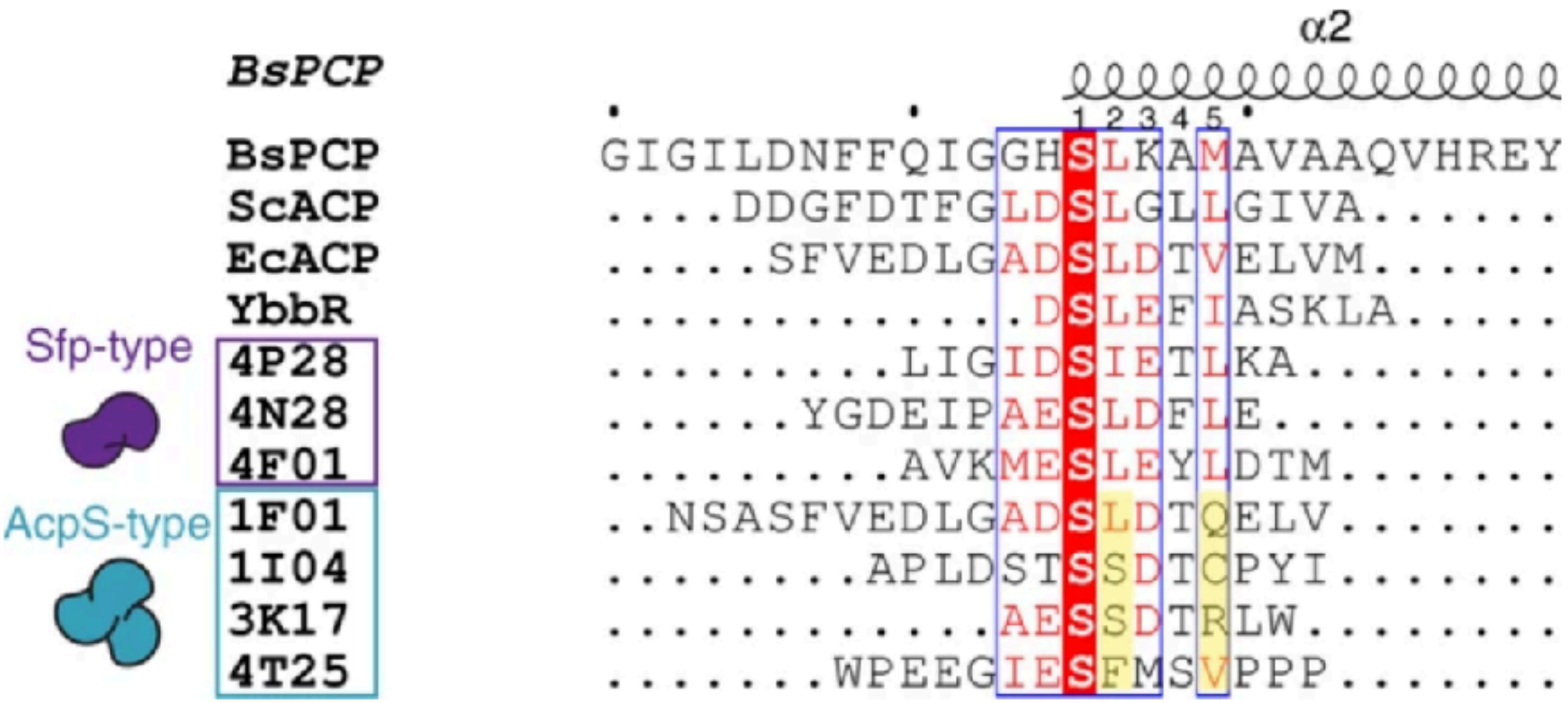


Summary

- POOL (peptide optimization with optimal learning) uses a BayesOpt-style approach to find short orthogonal peptide substrates.
- 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
- This method has found hits shorter than the shortest previously known.

Tallorin et al.“Discovering *de novo* peptide substrates for enzymes using machine learning.” *Nature Communications*, 2018

Appendix



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(≥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:

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

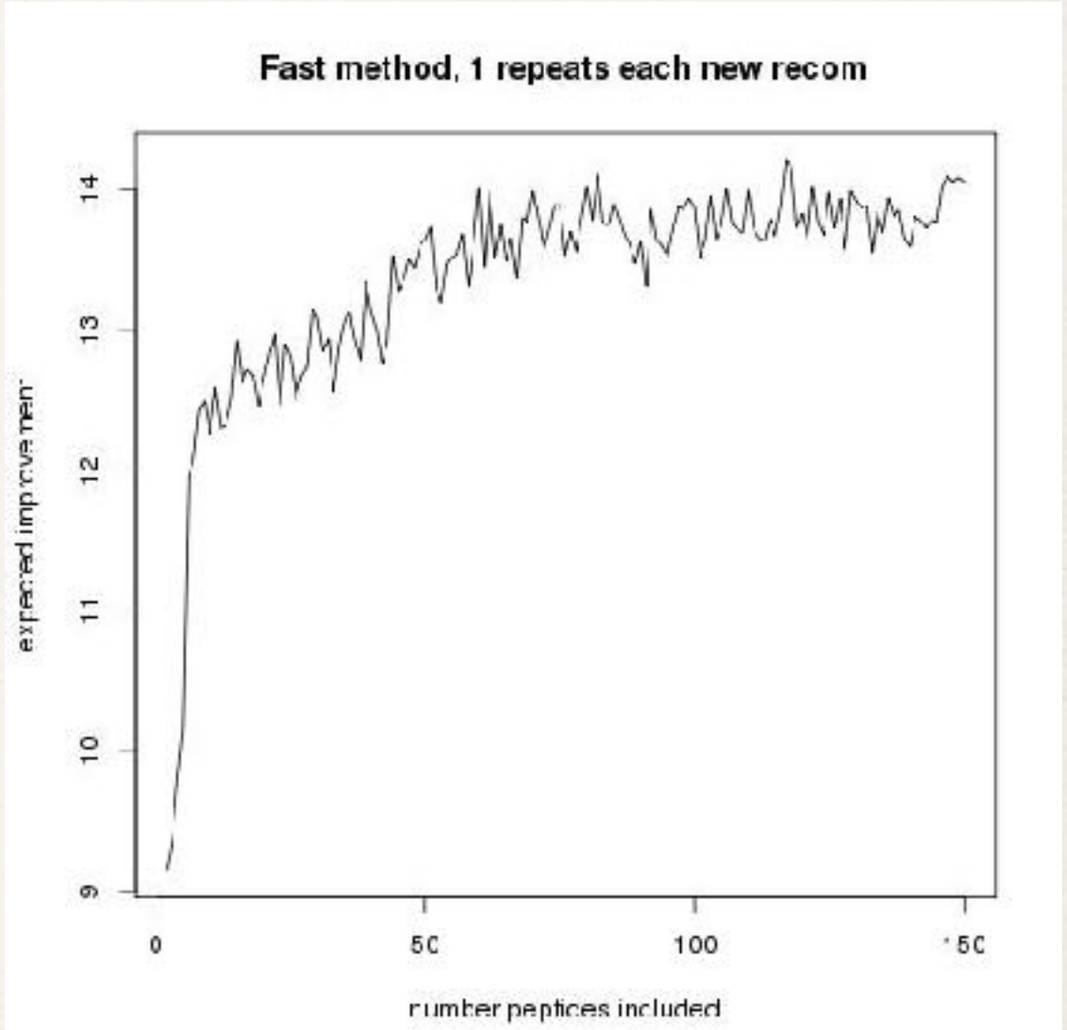
- An S that maximizes EI(S) could contain peptides shorter than b-1.

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.