

Optimal Learning for Peptide Design

Natural Materials, Systems and Extremophiles
Program Review
December 9-13, 2013

Warren Powell (PI), Princeton University
Peter Frazier (Co-PI), Cornell University

Supporting grant:
Optimal Learning for Efficient Experimentation in
Nanotechnology and Biochemistry





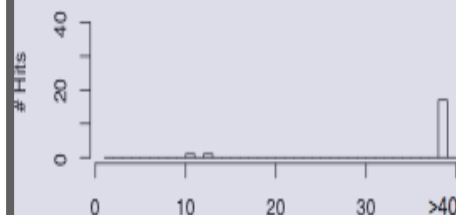
Optimal Learning for Peptide Design

Peter Frazier (Cornell), Warren Powell (Princeton),



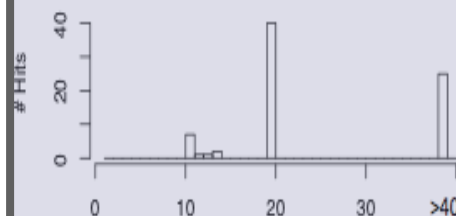
MAIN ACHIEVEMENTS

We have created a method, Peptide Optimization with Optimal Learning (**POOL**), that quickly finds minimal substrates for a pair of protein-modifying enzymes: PPTase and ACP hydrolase (with N. Gianneschi, M. Burkart, M. Gilson).



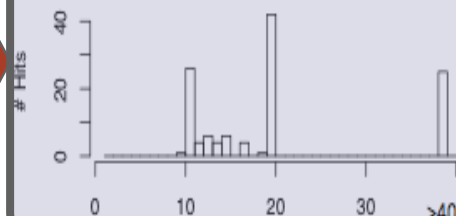
Training Set

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

Peptide Length

QUANTITATIVE IMPACT

POOL has revealed novel substrates **shorter than the previously best known**.

We are using POOL to create a system with orthogonal reactivity, providing **versatile building blocks** for bionanomaterials.

END-OF-PHASE GOAL

Our end goal is to provide an easy-to-use efficient method for designing peptides with any desired properties.

This will enable the creation of novel materials and catalytic complexes.

The capability to quickly design a peptide with desired properties would enable the creation of **novel materials**.

However, current approaches (ad hoc guessing; phage display) are **slow and unreliable**.

Machine learning can predict peptide properties.

Optimal learning can design experiments to quickly find desirable peptides.

Program goals

- Goal 1) Communicate the principles of optimal learning to the materials science community.
- Goal 2) Work with teams of scientists to help them apply optimal learning to accomplish their project goals.
- Goal 3) Develop new mathematical tools that meet challenges that arise in nano-bio research.
- Goal 3a) Develop new tools for belief extraction to capture the domain knowledge of scientists for an experiment.
 - Goal 3b) Expand knowledge-gradient to a wider range of belief models, motivated by settings that arise working with scientists in the nano-bio field.
 - Goal 3c) Develop other new tools as the needs of specific scientists arise.
- Goal 4) Implement an easy-to-use web-based tool that can be used by scientists for sequential design of experiments.

Goal progress assessment

Goal 1) Our ongoing work has allowed us to better understand the challenges that experimentalists encounter, and to frame our tools in the context of real problems.

Goal 2) We are working with six teams: the Prasad team based at Buffalo, the Mirkin group at Northwestern, the Gianneschi, Burkart and Gilson groups at UCSD, the Maruyama group at AFRL, the McAlpine group at Princeton, and the Clancy group at Cornell.

Goal 3) Progress includes:

- We have developed a new method for approximating the E max term of the knowledge gradient for nonlinear belief models, required to computing the value of information.
- We have developed a new optimal learning method for selecting peptides to test, in the search for minimal substrates for a pair of protein-modifying enzymes, and have used this method to find new shorter substrates. (in collaboration with Gianneschi/Burkart/Gilson).
- We have developed a statistical model for predicting contact residues and configurational entropy for peptide-gold binding (in collaboration with Walsh/Knecht/Prasad).

Goal 4) We have redesigned the web-based tool, now called Dr. hOLMES, for performing belief extraction and sequential experimental design. The new version is more intuitive. It should be available in the spring of 2014.

Transitions

- None to date.

Interactions with other groups

- We are working with the Gianneschi, Burkart and Gilson groups at UCSD to find short peptides that are minimized substrates for a pair of protein-modifying enzymes: phosphopantetheinyltransferase and ACP hydrolase.
- We are working with Paras Prasad (Buffalo) and his multi-university team funded by AFOSR (including Knecht and Walsh) on the development of 3D bio-mediated nanoparticle assembly paradigm for the production of reconfigurable biological nanoassemblies with useful photonic, electronic plasmonic and magnetic properties.
- We are working with the McAlpine group at Princeton to optimize surfactant concentrations to control the stability of nanoemulsions by identifying kinetic parameters through efficient experimentation.
- Working with Chad Mirkin's group, we have computed a knowledge gradient surface to guide find the best design of immobilized nanoparticles in a photoactive electrical device to maximize reflectivity.
- We are working with Paulette Clancy's group at Cornell apply optimal learning to design problems in organic photovoltaics.
- We are helping Benji Maruyama's group at AFRL on sequential policies to guide an experimental robot for optimizing carbon nanotubes.

Publications

- Appeared in 2013:
 - Lauren Hannah, W.B. Powell, D. Dunson, “Semi-Convex Regression for Metamodeling-Based Optimization,” SIAM J. on Optimization (to appear).
 - S. Dayanik, W.B. Powell, K. Yamazaki, “Asymptotically Optimal Bayesian Sequential Change Detection and Identification Rules,” Annals of Operations Research, (2013), Vol. 230, pp. 337-370. DOI 10.1007/s10479-012-1121-6
 - E. Barut and W. B. Powell, “Optimal Learning for Sequential Sampling with Non-Parametric Beliefs,” Journal of Global Optimization, DOI 10.1007/s10898-013-0050-5 (to appear)
 - Harvey Cheng, Arta Jamshidi, W. B. Powell, “The Knowledge Gradient Algorithm using Locally Parametric Approximations,” Winter Simulation Conference, Washington, D.C., 2013.
 - W. B. Powell, I. O. Ryzhov, “Optimal Learning and Approximate Dynamic Programming,” *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, (F. Lewis and D. Liu, eds.), John Wiley/CRC Press, New York, pp. 410-431, 2013.

Publications

- Appeared in 2013 (cont'd):
 - J. Xie, P.I. Frazier, “Sequential Bayes-Optimal Policies for Multiple Comparisons with a Known Standard,” *Operations Research*, vol. 61, no. 5, pp. 1174–1189, 2013.
 - [Winner, INFORMS Computing Society Student Paper Prize, 2013;
 - [Finalist, INFORMS Junior Faculty Interest Group (JFIG) Paper Competition, 2011]
 - R. Waeber, P.I. Frazier, S.G. Henderson, “Bisection Search with Noisy Responses,” *SIAM Journal on Control and Optimization*, vol. 51, no. 3, pp. 2261–2279, 2013.
 - L.H. Lee, E.P. Chew, P.I. Frazier, Q.S. Jia, and C.H. Chen “Advances in Simulation Optimization and its Applications” *IIE Transactions*, vol. 45, no. 7, pp. 683–684, 2013.
 - S.C. Clark, R. Egan, P.I. Frazier, and Z. Wang, “ALE: a Generic Assembly Likelihood Evaluation Framework for Assessing the Accuracy of Genome and Metagenome Assemblies,” *Bioinformatics*, vol. 29, no. 4, pp. 435–443, 2013.

Publications

- Appeared in 2013 (cont'd):
 - R. Sznitman, A. Lucchi, B. Jedynak, P.I. Frazier, P. Fua, “An Optimal Policy for Target Localization with Application to Electron Microscopy,” International Conference on Machine Learning (ICML), 2013.
 - A.J. Meltzer, A. Graham, P.H. Connolly, J.K. Karwowski, H.L. Bush, P.I. Frazier and D.B. Schneider, “Risk Factors for Early Failure after Peripheral Endovascular Intervention: Application of a Reliability Engineering Approach” Annals of Vascular Surgery, vol. 27, no. 1, pp. 53–61, 2013.
 - J. Xie, P.I. Frazier, “Upper Bounds for Bayesian Ranking & Selection” Winter Simulation Conference (WSC), 2013.

Publications

- In preparation/under review:
 - S. Chen, K. Reyes, M. Gupta, N. Masters, M. McAlpine and W.B. Powell, Adaptive learning in Experimental Design Using the Knowledge Gradient Policy with Application to Characterizing Nanoemulsion Stability
 - Yingfei Wang, K. Reyes, R. Boya, Q. Lin, K. Brown, C. Mirkin and W.B. Powell, “Nested batch learning for adaptive experimental design of DNA-functionalized nanoparticle photoactive devices”
 - Arta Jamshidi and W. B. Powell, “A Recursive Semi-parametric Approximation Method using Dirichlet Clouds and Radial Basis Functions,” (under second review at SIAM J. on Scientific Computing).
 - Boris Defourny, Ilya O. Ryzhov, W. B. Powell, “Optimal Information Blending with Measurements in the L2 Sphere,” (under revision for resubmission to Mathematics of Operations Research, October 12, 2012.)
 - Samuel J. Gershman, P.I. Frazier, David M. Blei, “Distance Dependent Infinite Latent Feature Models,” in second review IEEE Trans. Pattern Analysis and Machine Intelligence.

Publications

- In preparation/under review:
 - P.I. Frazier, “A Fully Sequential Elimination Procedure for Indifference-Zone Ranking and Selection with Tight Bounds on Probability of Correct Selection,” in review at Operations Research.
 - [Finalist, INFORMS Junior Faculty Interest Group (JFIG) Paper Competition, 2013]
 - I.O. Ryzhov, P.I. Frazier, and W.B. Powell, “A New Optimal Stepsize Rule for Approximate Dynamic Programming,” in review at IEEE Transactions on Automatic Control.
 - J. Xie, P.I. Frazier, and S.E. Chick, “Bayesian Optimization via Simulation with Pairwise Sampling and Correlated Prior Beliefs.” in review at Operations Research.
 - P.I. Frazier, M. Knecht, P. Palafox-Hernandez, T.R. Walsh, J. Wang, “Optimal Learning for Peptide Design”, in preparation.
 - M. Burkart, P.I. Frazier, N. Gianneschi, M. Gilson, N. Kosa, M. Rothmann, L. Tallorin, J. Wang, P. Yang, “An Active Learning Approach to Finding Minimally-sized Peptide Substrates”, in preparation.

Transformational/evolutionary

- Our research has the potential for transforming the fundamental way in which research is conducted in the physical sciences, providing a set of rigorous tools that make experimental success in high-risk high-reward settings more attainable.
- Our work goes well beyond classical methods of experimental design, which are less efficient and less able to incorporate experimentalists' intuition and domain expertise.
- Our work will help scientists estimate the likelihood of success in experiments with a high level of uncertainty, helping them assess risks and rewards from different experimental strategies.

We apply optimal learning to two peptide-design problems

1. Finding minimized substrates for a pair of protein-modifying enzymes.

– Joint work with Nathan Gianneschi, Michael Burkart, Michael Gilson

2. Finding specific binders for a given pair of target materials

– Joint work with Tiff Walsh and Marc Knecht, working with Paras Prasad, Mark Swihart, and Aidong Zhang.

We will focus on the first problem in this talk

1. Finding minimized substrates for a pair of protein-modifying enzymes.

– Joint work with Nathan Gianneschi, Michael Burkart, Michael Gilson


2. Finding specific binders for a given pair of target materials

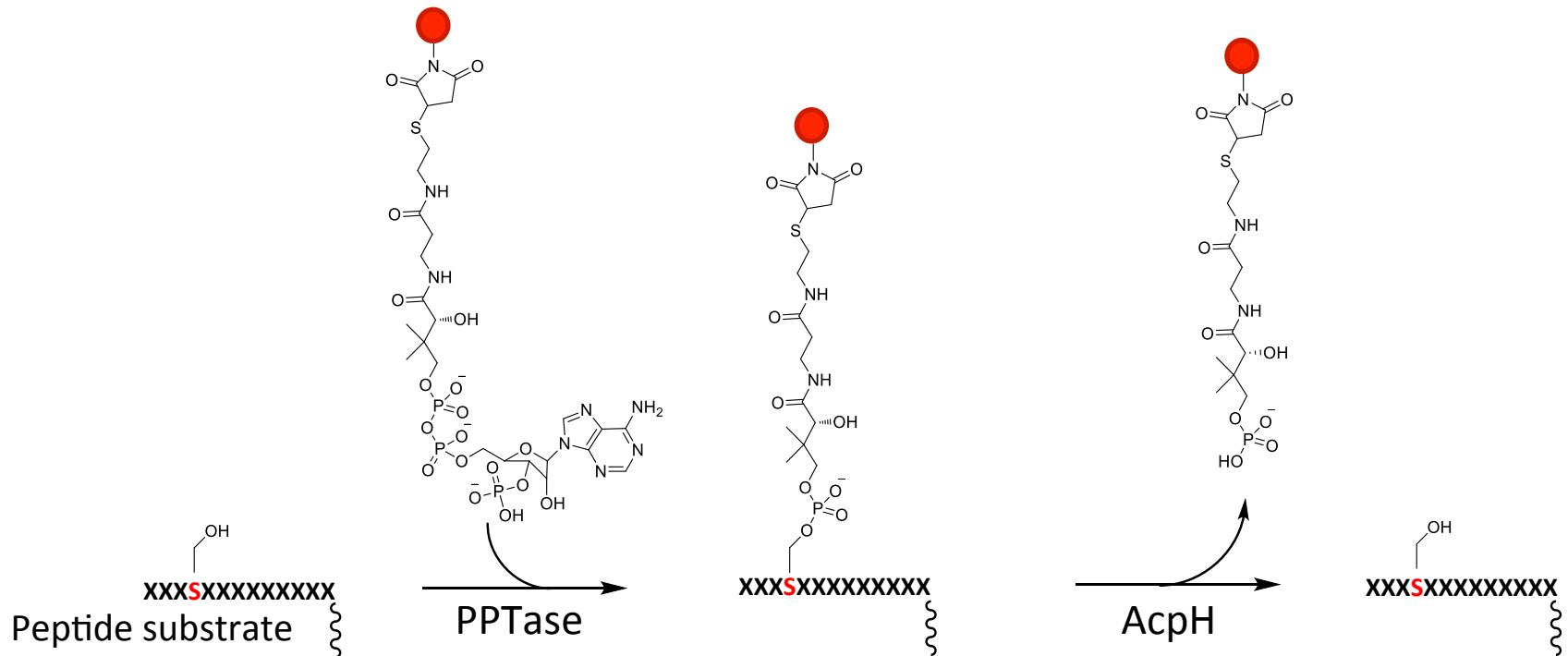
– Joint work with Tiff Walsh and Marc Knecht, working with Paras Prasad, Mark Swihart, and Aidong Zhang.

We will focus on the first problem in this talk

1. Finding minimized substrates for a pair of protein-modifying enzymes.
 - Joint work with Nathan Gianneschi, Michael Burkart, Michael Gilson,
 - And also Nick Kosa, Michael Rothmann, Lorillee Tallorin, Jialei Wang, Pu Yang

A minimized substrate would enable versatile building blocks for biomaterials

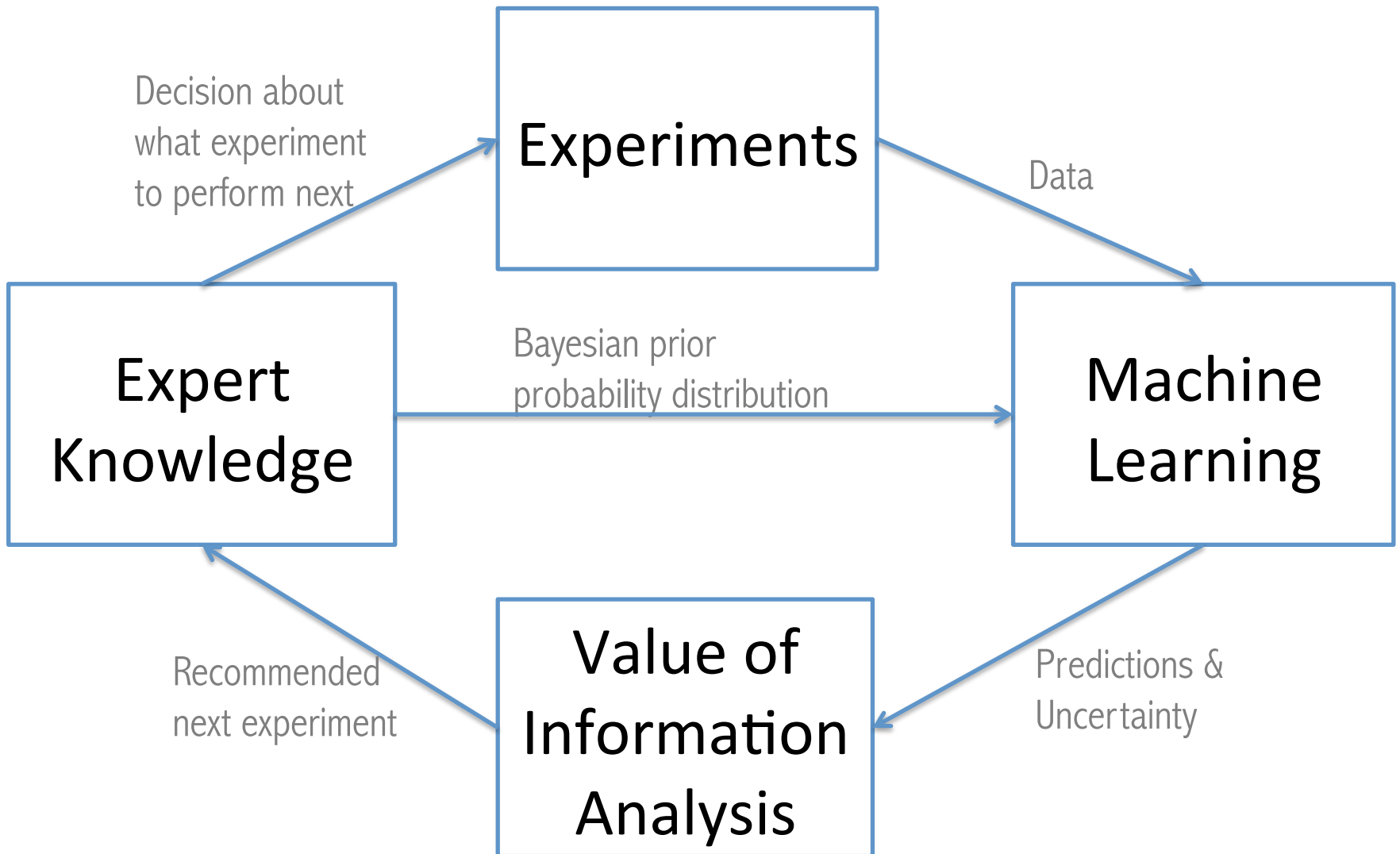
- We consider two protein-modifying enzymes:
 - PPTase: attaches an arbitrary label to the peptide substrate,  = misc label at a conserved serine.
 - AcpH: removes the label



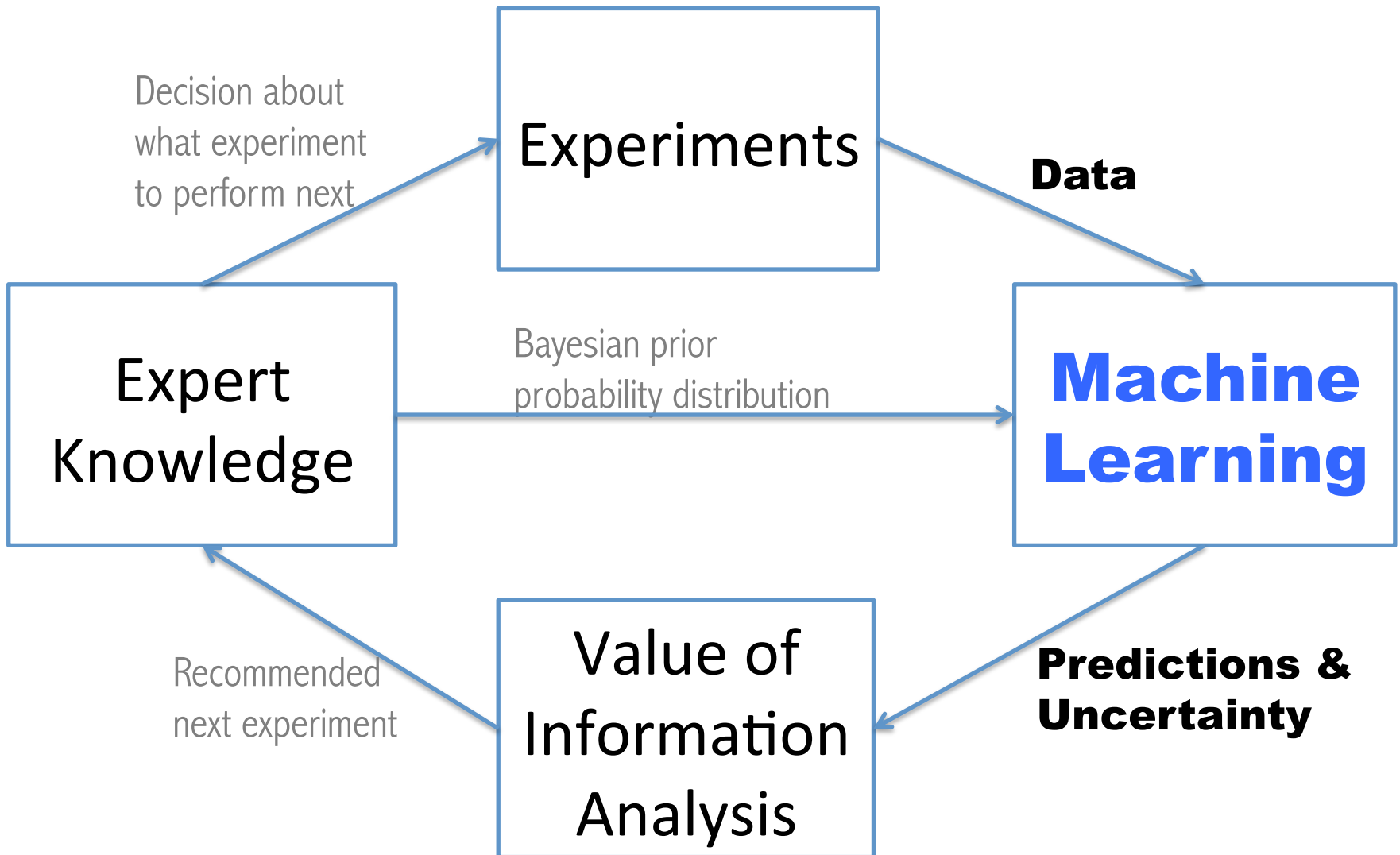
We reduce the experimental effort required to find minimal substrates

- We provide a method for Peptide Optimization with Optimal Learning (**POOL**).
- Our method has two parts:
 - Predict which peptides are “hits”.
 - Based on these predictions, recommend which peptides to test next.

Peptide Optimization with Optimal Learning (**POOL**)



First, we consider prediction.



We use a machine learning method to predict which peptides will be hits

- We use a method called “Naïve Bayes.”
 - It is called “naïve” because it makes an independence assumption.
 - Even though it makes this assumption, it often works quite well.
- This method is commonly used for text classification, e.g., for spam filtering.
- We have adapted it to peptide prediction.

Here's how Naïve Bayes Works

- Pick a particular position relative to the conserved serine, say +3, and a particular amino acid, A.
- Imagine we draw a peptide at random from the set of hits
 - (we can't, but imagine we could).
- Imagine we do this many times, calculate the fraction of time the hit has an A at this position, and put it in the table below.

[illegible]

Here's how Naïve Bayes Works

- If we knew these two tables, call them HIT and MISS, and we had a rough guess at the fraction p of peptides that are hits...
- Then, given a random peptide x , we could calculate the probability it is a hit using Bayes rule as:

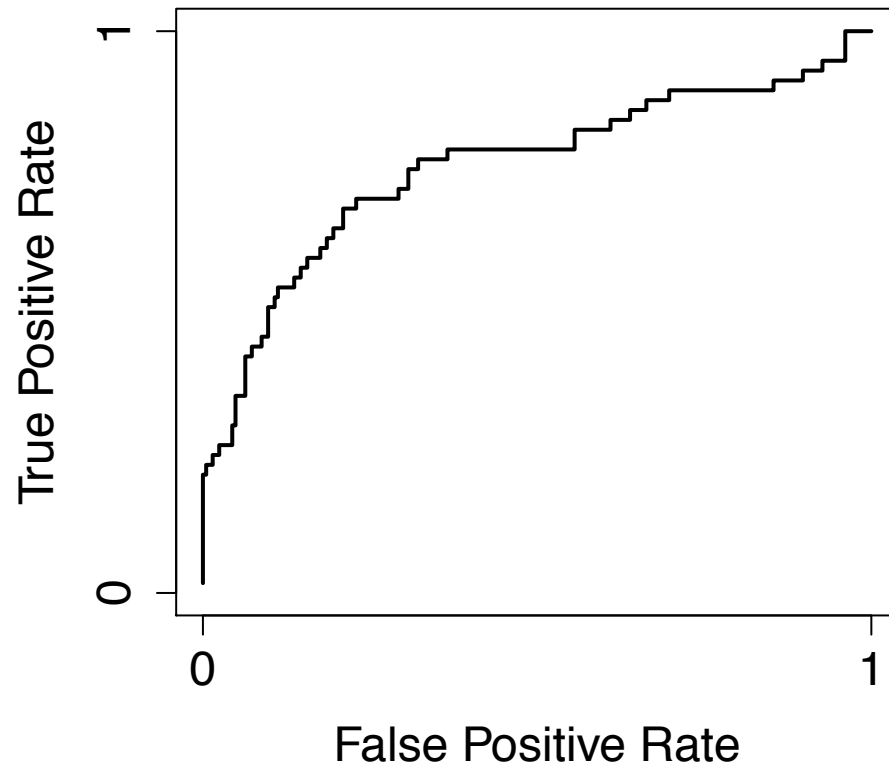
$$P(x \text{ is a hit}) = \frac{p \prod_i \text{HIT}(i, x_i)}{p \prod_i \text{HIT}(i, x_i) + (1 - p) \prod_i \text{MISS}(i, x_i)}$$

Here's how Naïve Bayes Works

- To use this method, we just need a way to estimate the HIT and MISS tables.
- If we had unlimited data, we could simply use the empirical distribution.
 - i.e., fill in the fraction of hits with an A at position +3 at the corresponding location in the table.
- Since our data is limited, we also use domain expertise to improve performance:
 - We use a reduced amino acid alphabet
 - We use a Bayesian prior distribution
 - MISS's values are likely to be close to 5%
 - HIT's values are likely to vary near the serene.
 - HIT's values are likely to be close to 5% far from the serene.

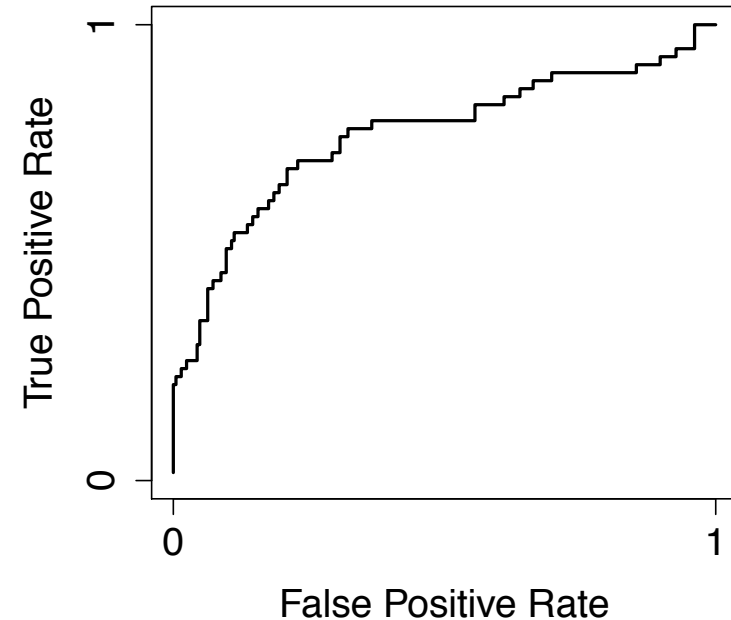
This ROC curve shows that Naïve Bayes predicts well

This ROC curve is computed using leave-one-out cross validation, on a training set of 262 peptides.



Review of ROC curves

Peptide	Predicted Probability of a Hit	Prediction	Reality
DALEFIASKLA	.729		Hit
MGICSSSSWIYGME	.721		Hit
DSAEFIASKLA	.719		Hit
ASLEFIASKLA	.709		Hit
VSMESSETLMLPIE	.690		Miss
DSLEFIAAKLA	.620		Hit
FGLDSTSSIVVSAE	.585		Miss
ADSTETMMMTSE	.340		Hit
YPIDSTDTGVMSVD	.295		Miss
...

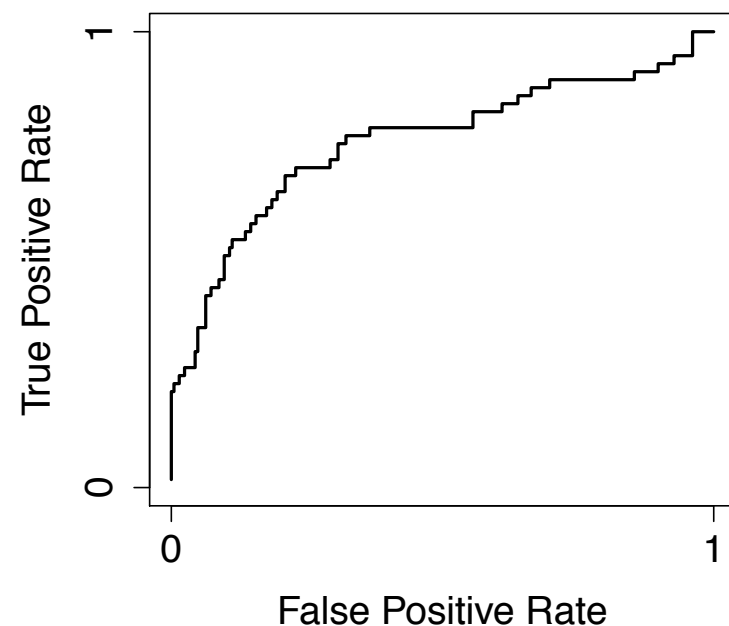


True Positive Rate =
 $\frac{\# \text{ true hits predicted as hits}}{\# \text{ true hits}}$.

False Positive Rate =
 $\frac{\# \text{ true misses predicted as hits}}{\# \text{ true misses}}$.

Review of ROC curves

Peptide	Predicted Probability of a Hit	Prediction	Reality
DALEFIASKLA	.729	Hit	Hit
MGICSSSSWIYGME	.721	Miss	Hit
DSAEFIASKLA	.719	Miss	Hit
ASLEFIASKLA	.709	Miss	Hit
VSMESSETLMLPIE	.690	Miss	Miss
DSLEFIAAKLA	.620	Miss	Hit
FGLDSTSSIVVSAE	.585	Miss	Miss
ADSTETMMMTSE	.340	Miss	Hit
YPIDSTDTGVMSVD	.295	Miss	Miss
...

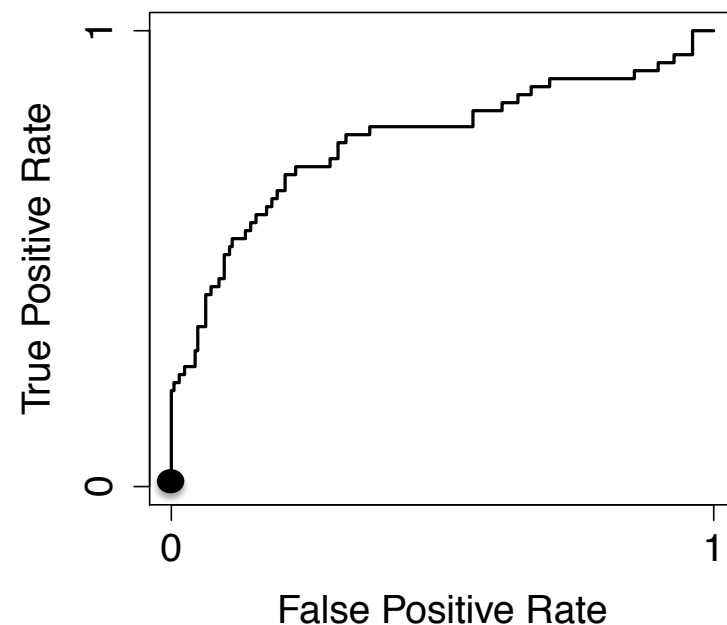


True Positive Rate =
 $\frac{\# \text{ true hits predicted as hits}}{\# \text{ true hits.}}$

False Positive Rate =
 $\frac{\# \text{ true misses predicted as hits}}{\# \text{ true misses.}}$

Review of ROC curves

Peptide	Predicted Probability of a Hit	Prediction	Reality
DALEFIASKLA	.729	Hit	Hit
MGICSSSSWIYGME	.721	Miss	Hit
DSAEFIASKLA	.719	Miss	Hit
ASLEFIASKLA	.709	Miss	Hit
VSMESSETLMLPIE	.690	Miss	Miss
DSLEFIAAKLA	.620	Miss	Hit
FGLDSTSSIVVSAE	.585	Miss	Miss
ADSTETMMMTSE	.340	Miss	Hit
YPIDSTDTGVMSVD	.295	Miss	Miss
...

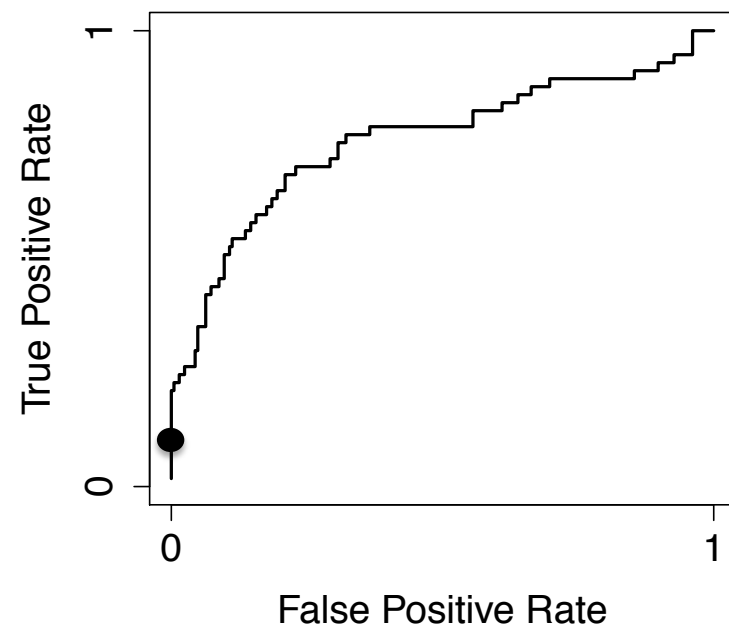


True Positive Rate =
 $\frac{\text{\# true hits predicted as hits}}{\text{\# true hits}}$

False Positive Rate =
 $\frac{\text{\# true misses predicted as hits}}{\text{\# true misses}}$

Review of ROC curves

Peptide	Predicted Probability of a Hit	Prediction	Reality
DALEFIASKLA	.729	Hit	Hit
MGICSSSSWIYGME	.721	Hit	Hit
DSAEFIASKLA	.719	Miss	Hit
ASLEFIASKLA	.709	Miss	Hit
VSMESSETLMLPIE	.690	Miss	Miss
DSLEFIAAKLA	.620	Miss	Hit
FGLDSTSSIVVSAE	.585	Miss	Miss
ADSTETMMMTSE	.340	Miss	Hit
YPIDSTDTGVMSVD	.295	Miss	Miss
...

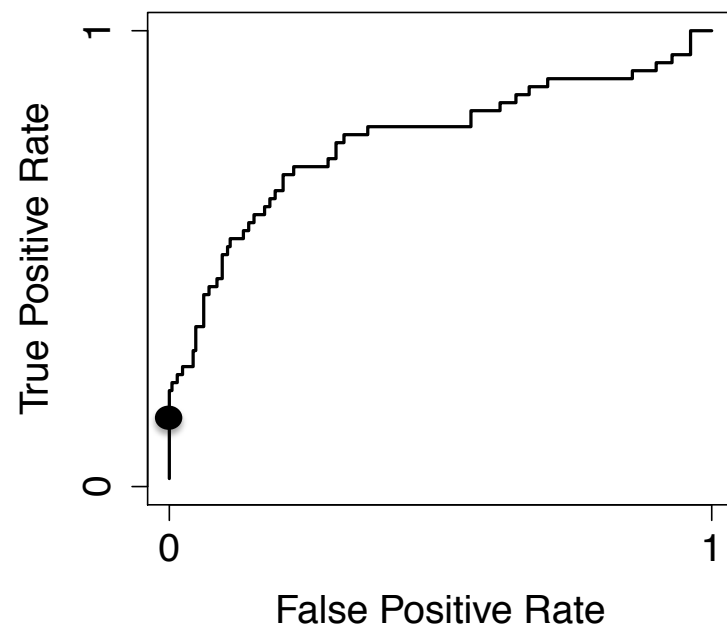


True Positive Rate =
 $\frac{\text{\# true hits predicted as hits}}{\text{\# true hits}}$

False Positive Rate =
 $\frac{\text{\# true misses predicted as hits}}{\text{\# true misses}}$

Review of ROC curves

Peptide	Predicted Probability of a Hit	Prediction	Reality
DALEFIASKLA	.729	Hit	Hit
MGICSSSSWIYGME	.721	Hit	Hit
DSAEFIASKLA	.719	Hit	Hit
ASLEFIASKLA	.709	Miss	Hit
VSMESSETLMLPIE	.690	Miss	Miss
DSLEFIAAKLA	.620	Miss	Hit
FGLDSTSSIVVSAE	.585	Miss	Miss
ADSTETMMMTSE	.340	Miss	Hit
YPIDSTDTGVMSVD	.295	Miss	Miss
...

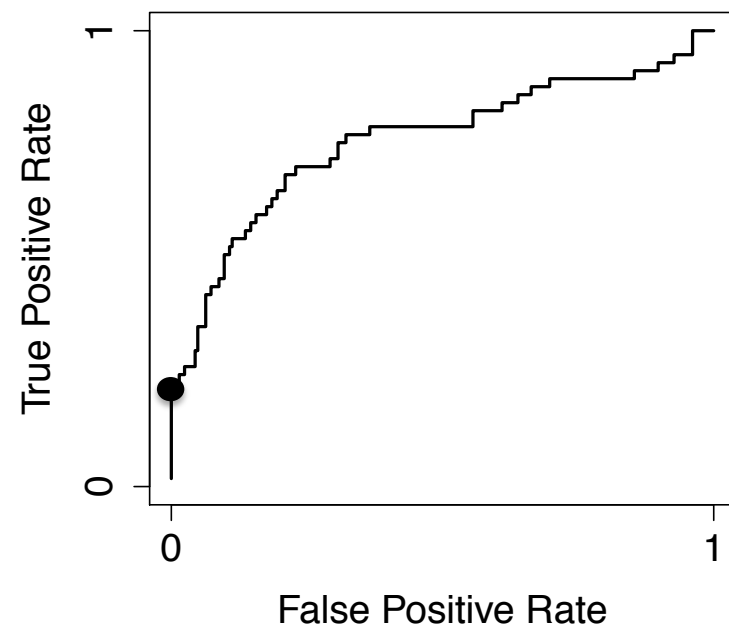


True Positive Rate =
 $\frac{\text{\# true hits predicted as hits}}{\text{\# true hits}}$

False Positive Rate =
 $\frac{\text{\# true misses predicted as hits}}{\text{\# true misses}}$

Review of ROC curves

Peptide	Predicted Probability of a Hit	Prediction	Reality
DALEFIASKLA	.729	Hit	Hit
MGICSSSSWIYGME	.721	Hit	Hit
DSAEFIASKLA	.719	Hit	Hit
ASLEFIASKLA	.709	Hit	Hit
VSMESSETLMLPIE	.690	Miss	Miss
DSLEFIAAKLA	.620	Miss	Hit
FGLDSTSSIVVSAE	.585	Miss	Miss
ADSTETMMMTSE	.340	Miss	Hit
YPIDSTDTGVMSVD	.295	Miss	Miss
...

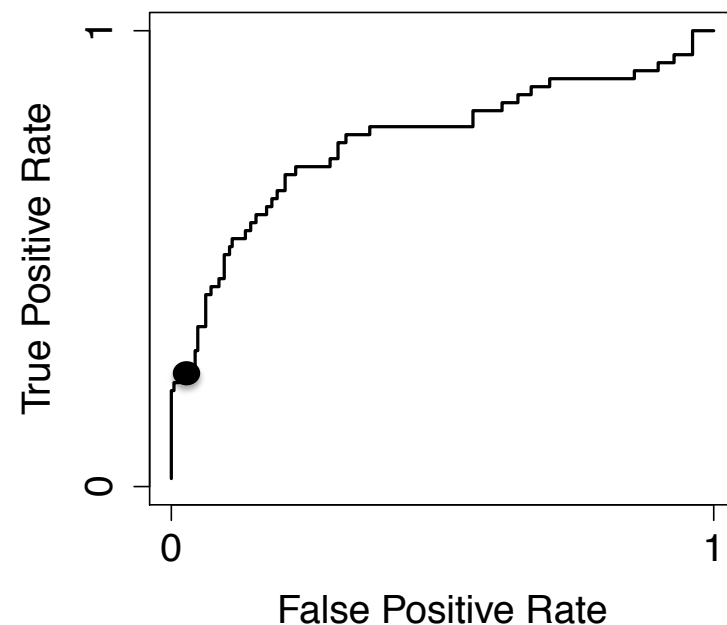


True Positive Rate =
 $\frac{\text{\# true hits predicted as hits}}{\text{\# true hits}}$

False Positive Rate =
 $\frac{\text{\# true misses predicted as hits}}{\text{\# true misses}}$

Review of ROC curves

Peptide	Predicted Probability of a Hit	Prediction	Reality
DALEFIASKLA	.729	Hit	Hit
MGICSSSSWIYGME	.721	Hit	Hit
DSAEFIASKLA	.719	Hit	Hit
ASLEFIASKLA	.709	Hit	Hit
VSMESSETLMLPIE	.690	Hit	Miss
DSLEFIAAKLA	.620	Miss	Hit
FGLDSTSSIVVSAE	.585	Miss	Miss
ADSTETMMMTSE	.340	Miss	Hit
YPIDSTDTGVMSVD	.295	Miss	Miss
...

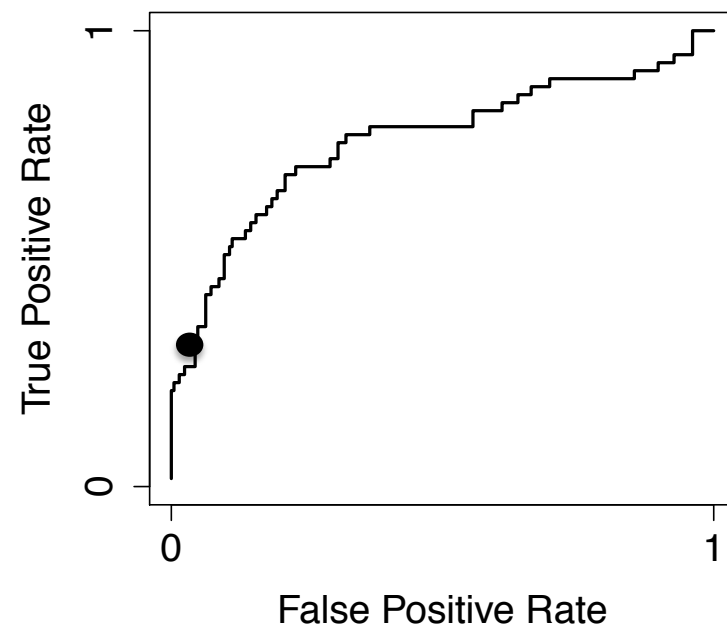


True Positive Rate =
 $\frac{\text{\# true hits predicted as hits}}{\text{\# true hits}}$

False Positive Rate =
 $\frac{\text{\# true misses predicted as hits}}{\text{\# true misses}}$

Review of ROC curves

Peptide	Predicted Probability of a Hit	Prediction	Reality
DALEFIASKLA	.729	Hit	Hit
MGICSSSSWIYGME	.721	Hit	Hit
DSAEFIASKLA	.719	Hit	Hit
ASLEFIASKLA	.709	Hit	Hit
VSMESSETLMLPIE	.690	Hit	Miss
DSLEFIAAKLA	.620	Hit	Hit
FGLDSTSSIVVSAE	.585	Miss	Miss
ADSTETMMMTSE	.340	Miss	Hit
YPIDSTDTGVMSVD	.295	Miss	Miss
...

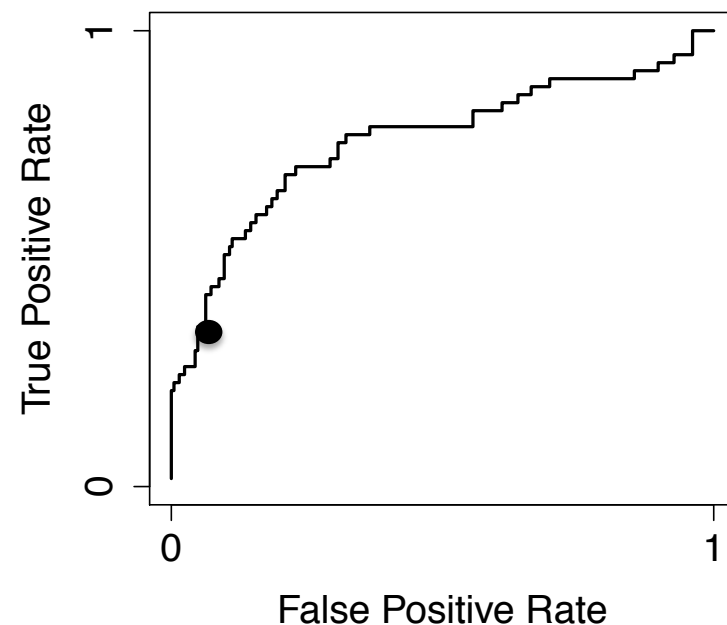


True Positive Rate =
 $\frac{\# \text{ true hits predicted as hits}}{\# \text{ true hits}}$

False Positive Rate =
 $\frac{\# \text{ true misses predicted as hits}}{\# \text{ true misses}}$

Review of ROC curves

Peptide	Predicted Probability of a Hit	Prediction	Reality
DALEFIASKLA	.729	Hit	Hit
MGICSSSSWIYGME	.721	Hit	Hit
DSAEFIASKLA	.719	Hit	Hit
ASLEFIASKLA	.709	Hit	Hit
VSMESSETLMLPIE	.690	Hit	Miss
DSLEFIAAKLA	.620	Hit	Hit
FGLDSTSSIVVSAE	.585	Hit	Miss
ADSTETMMMTSE	.340	Miss	Hit
YPIDSTDTGVMSVD	.295	Miss	Miss
...

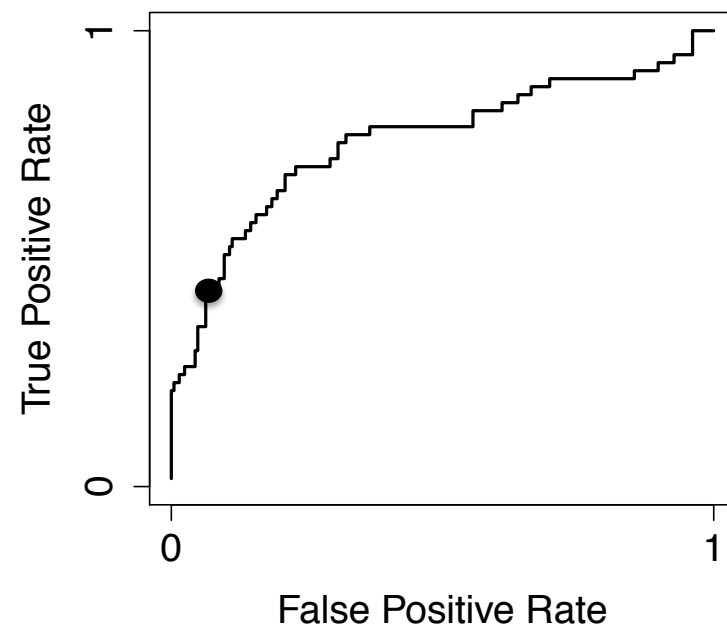


True Positive Rate =
 $\frac{\text{\# true hits predicted as hits}}{\text{\# true hits}}$

False Positive Rate =
 $\frac{\text{\# true misses predicted as hits}}{\text{\# true misses}}$

Review of ROC curves

Peptide	Predicted Probability of a Hit	Prediction	Reality
DALEFIASKLA	.729	Hit	Hit
MGICSSSSWIYGME	.721	Hit	Hit
DSAEFIASKLA	.719	Hit	Hit
ASLEFIASKLA	.709	Hit	Hit
VSMESSETLMLPIE	.690	Hit	Miss
DSLEFIAAKLA	.620	Hit	Hit
FGLDSTSSIVVSAE	.585	Hit	Miss
ADSTETMMMTSE	.340	Hit	Hit
YPIDSTDTGVMSVD	.295	Miss	Miss
...



True Positive Rate =
 $\frac{\text{\# true hits predicted as hits}}{\text{\# true hits}}$

False Positive Rate =
 $\frac{\text{\# true misses predicted as hits}}{\text{\# true misses}}$

Review of ROC curves

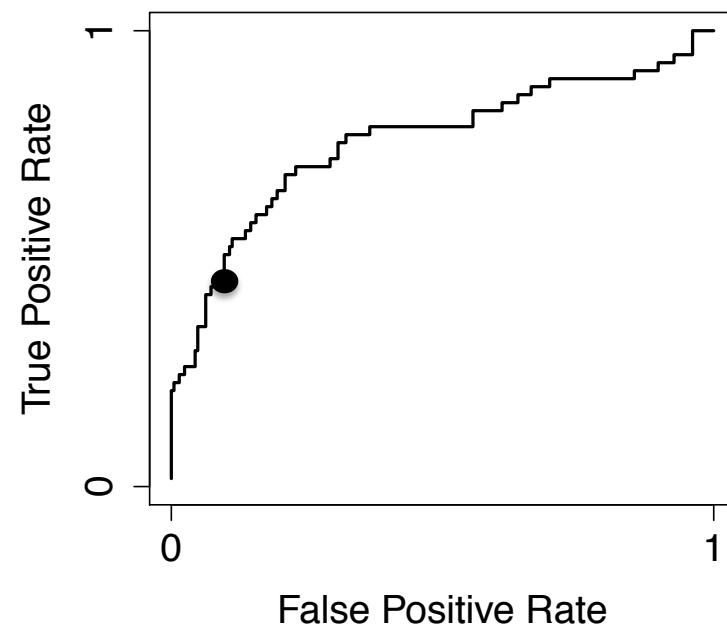
Peptide	Predicted Probability of a Hit	Prediction	Reality
DALEFIASKLA	.729	Hit	Hit
MGICSSSSWIYGME	.721	Hit	Hit
DSAEFIASKLA	.719	Hit	Hit
ASLEFIASKLA	.709	Hit	Hit
VSMESSETLMLPIE	.690	Hit	Miss
DSLEFIAAKLA	.620	Hit	Hit
FGLDSTSSIVVSAE	.585	Hit	Miss
ADSTETMMMTSE	.340	Hit	Hit
YPIDSTDTGVMSVD	.295	Hit	Miss

...

...

...

...

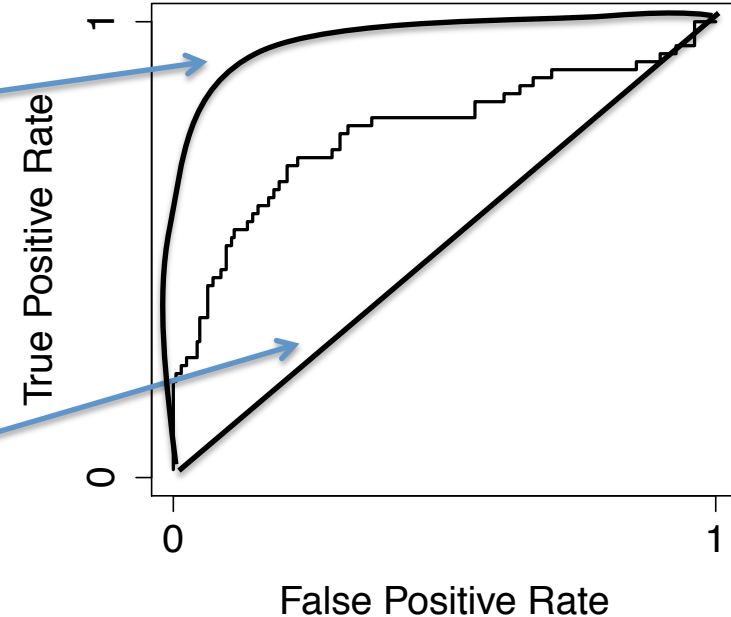


True Positive Rate =
 $\frac{\text{\# true hits predicted as hits}}{\text{\# true hits}}$

False Positive Rate =
 $\frac{\text{\# true misses predicted as hits}}{\text{\# true misses}}$

Review of ROC curves

- Better prediction methods have ROC curves further up and to the left.
- Predicting probability of a hit using a random number generator will give a diagonal line.

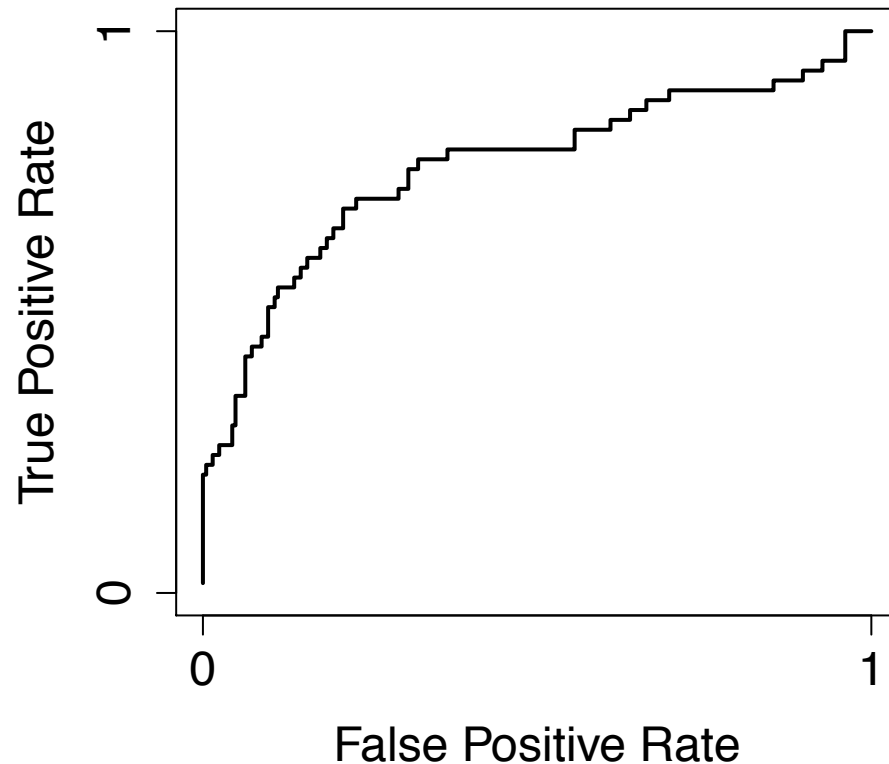


True Positive Rate =
 $\frac{\text{\# true hits predicted as hits}}{\text{\# true hits}}$

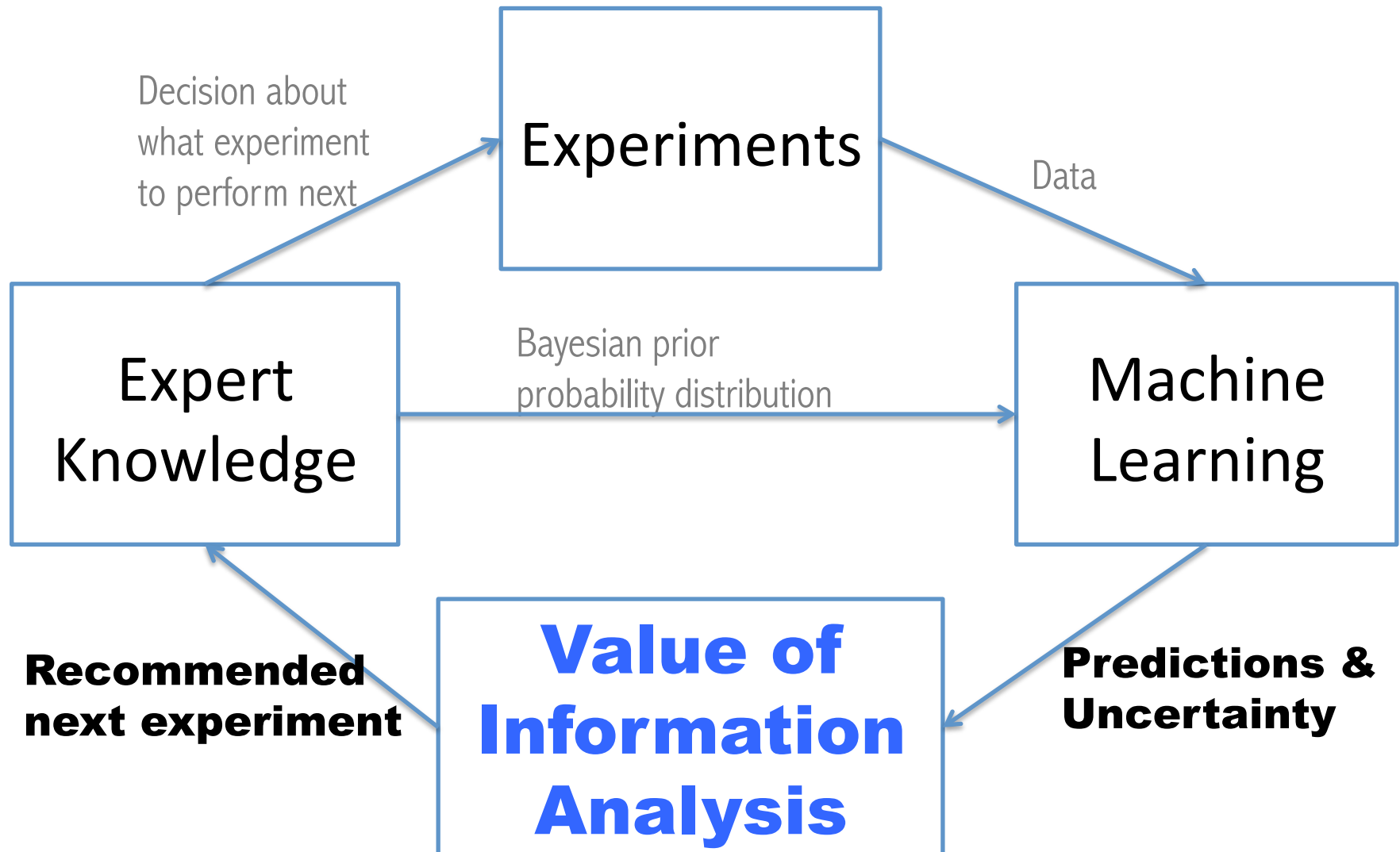
False Positive Rate =
 $\frac{\text{\# true misses predicted as hits}}{\text{\# true misses}}$

This ROC curve shows that Naïve Bayes predicts well

This ROC curve is computed using leave-one-out cross validation, on a training set of 262 peptides.

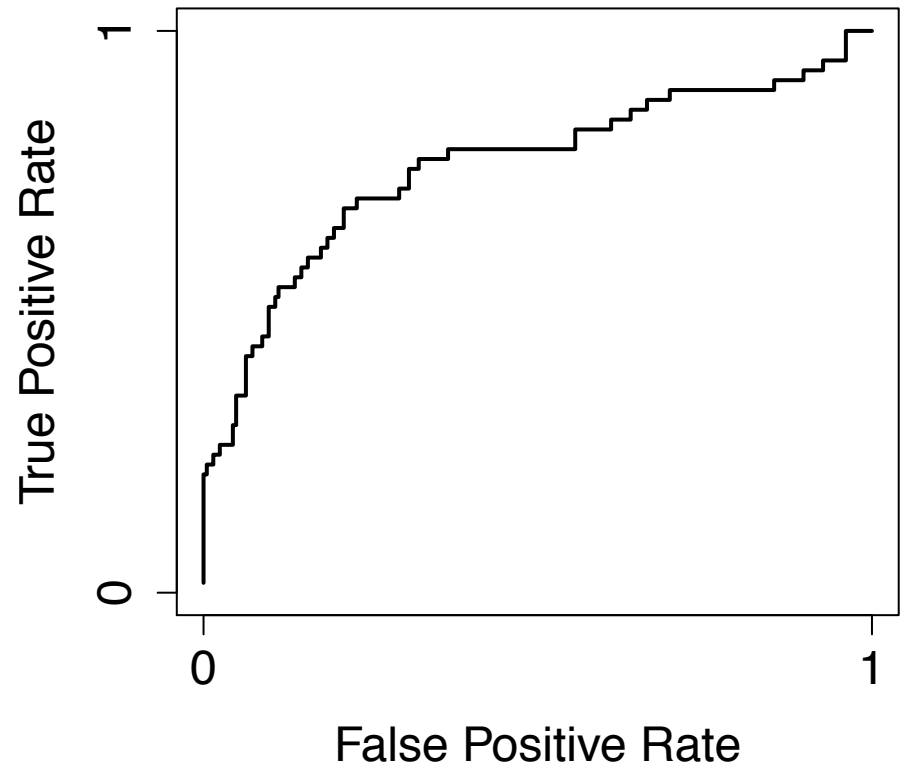


Now, we consider the choice of experiment.



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?

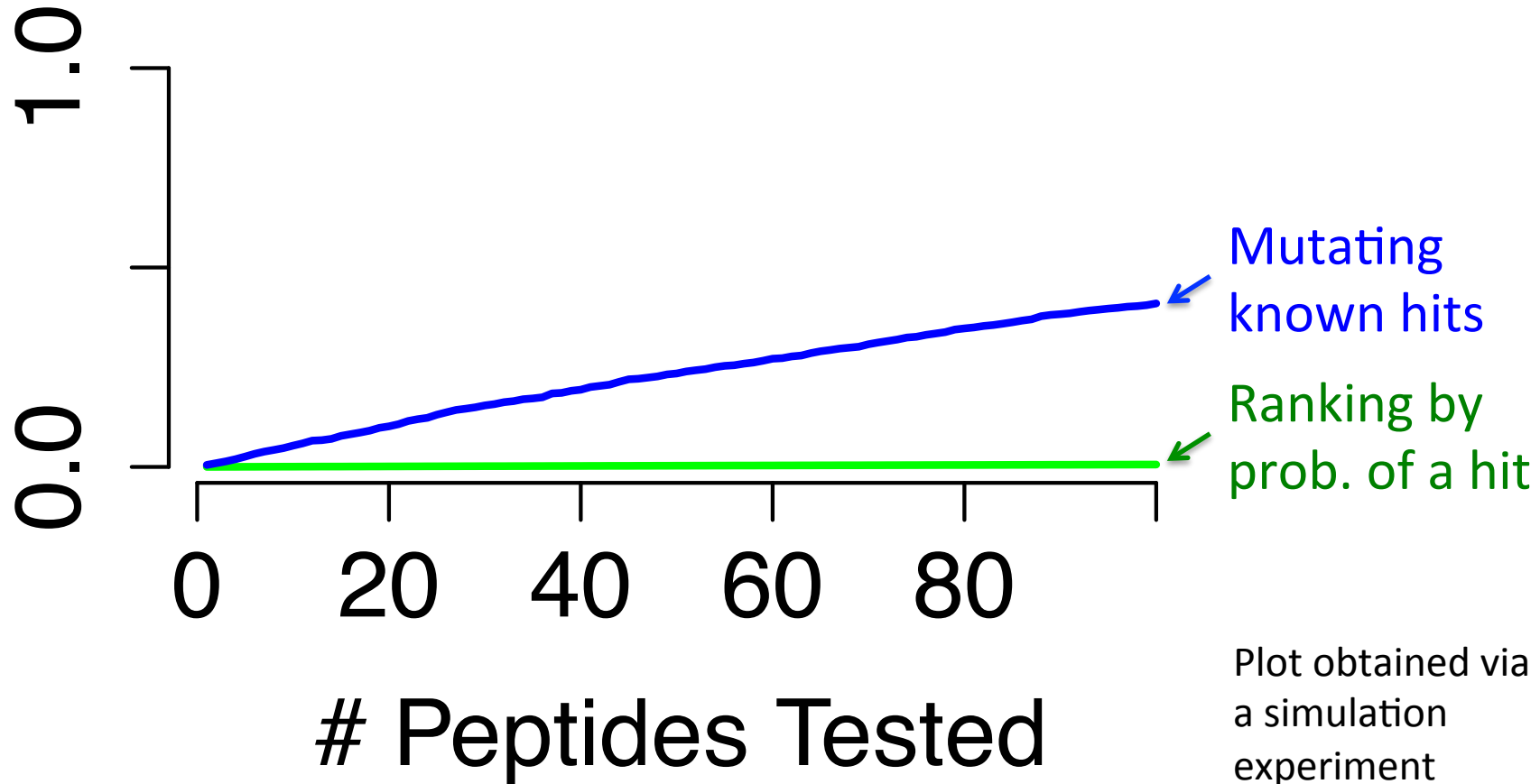


Ranking by probability of a hit does not work well

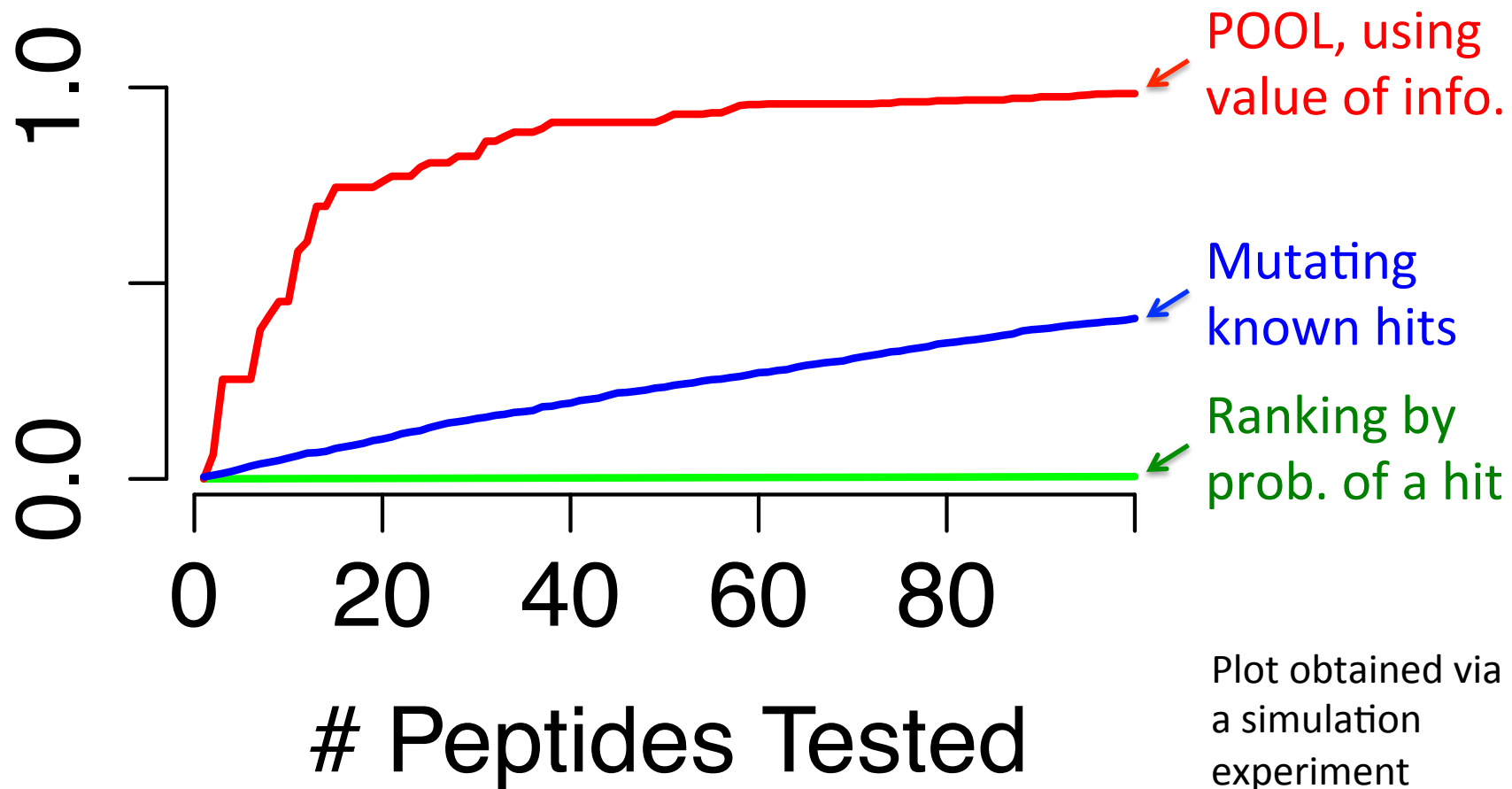
- One simple strategy is:
 - Select those peptides with length < 12 .
 - Rank those peptides by predicted probability of a hit
 - Test the top 300 most likely to be a hit.
- The tested peptides are very similar. If the first tested peptide is not a hit, the other ones probably aren't either.

Ranking by probability of a hit does not work well

Probability of a short hit



We can do better using value of information (VOI)



VOI says: choose the experiment that maximizes the probability of reaching our goal

- Our goal is to find short hits.
- More specifically:
 - Our goal is to find at least one hit shorter than a target length b . For example, we could set $b=10$.
- We should then choose the experiment to run that maximizes the probability of reaching this goal.

VOI says: choose the experiment that maximizes the probability of reaching our goal

- Mathematically, this can be formulated as this optimization problem:

$$\max_{S \subseteq E: |S|=300} P(\text{at least one short hit in } S)$$

- Notation:
 - E is the set of all peptides.
 - S is the set of peptides to test.
 - S is required to have less than 300 peptides so it can be tested in a single experiment.
 - A “short hit” is a hit whose length is less than b.
- Solving this optimization problem exactly is difficult. We use a greedy approach to obtain an approximate solution.

We find the best experiment using a greedy approach

- We build up the set S of peptides to test in stages.
- In each stage, we find the single peptide to add that maximizes the probability of reaching our goal:

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

- We then add e to S and repeat, until S has 300 peptides.

There is a reason why VOI works better

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

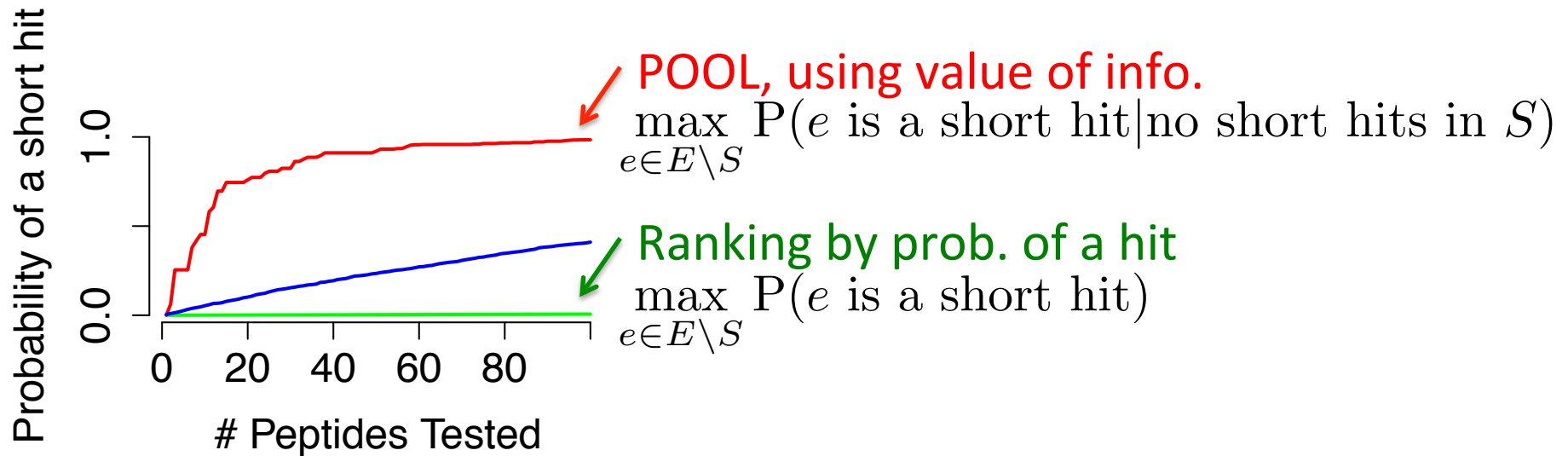
$$\max_{e \in E \setminus S} P(\text{at least one short hit in } S \cup \{e\})$$

- Is equivalent to:

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

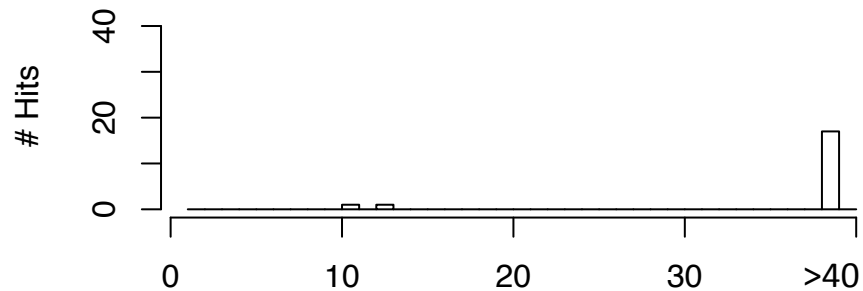
- Compare this to the “rank by prob. hit” approach
$$\max_{e \in E \setminus S} P(e \text{ is a short hit})$$

VOI works better because its peptides are more diverse



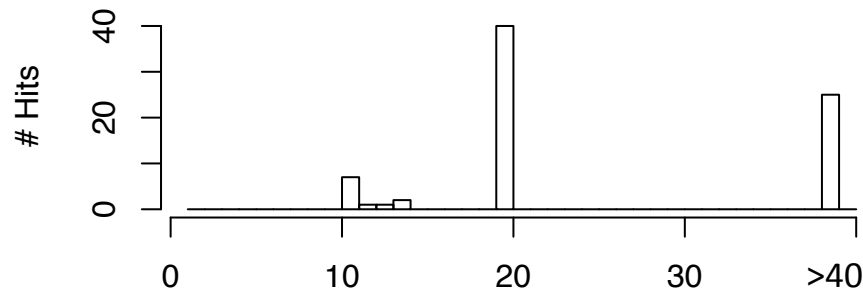
- Peptides added using the value of information approach tend to be **different** from those already in S .
- Its recommendations are more **diverse**.

We have found novel short peptides using this method



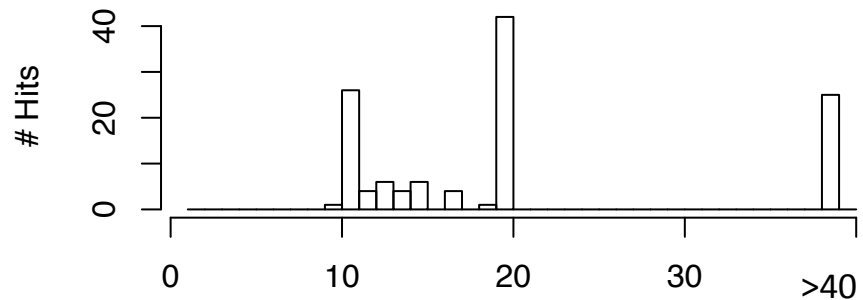
Training Set

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

Peptide Length

Ongoing: create a system with orthogonal reactivity

- Given two PPTase homologues, **PPTase 1** and **PPTase 2**, and one **AcpH**, we will design two minimal peptides substrates:
 - **Peptide 1** will be a substrate for **PPTase 1** and **AcpH**, but not **PPTase 2**.
 - **Peptide 2** will be a substrate for **PPTase 2** and **AcpH**, but not **PPTase 1**.
- POOL can design each of these peptides, simply by redefining what it means to be a “hit”.

Ongoing: create a system with orthogonal reactivity

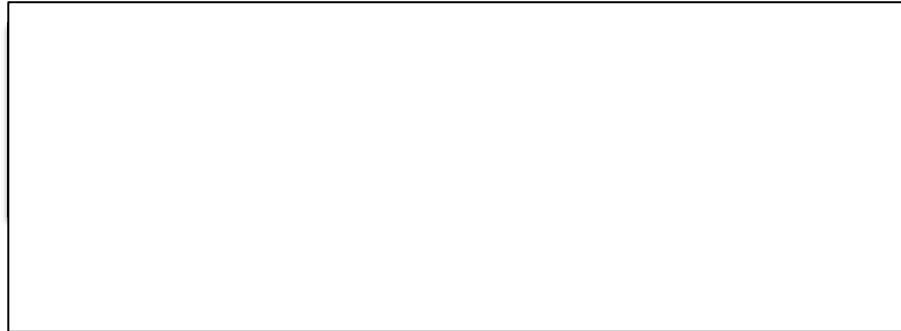
Put **Peptide 1** in the letters



Put **Peptide 2** in the blank space

Ongoing: create a system with orthogonal reactivity

The slide begins blank



Ongoing: create a system with orthogonal reactivity

Add PPTase 1
with a blue label



AFOSR

Ongoing: create a system with orthogonal reactivity

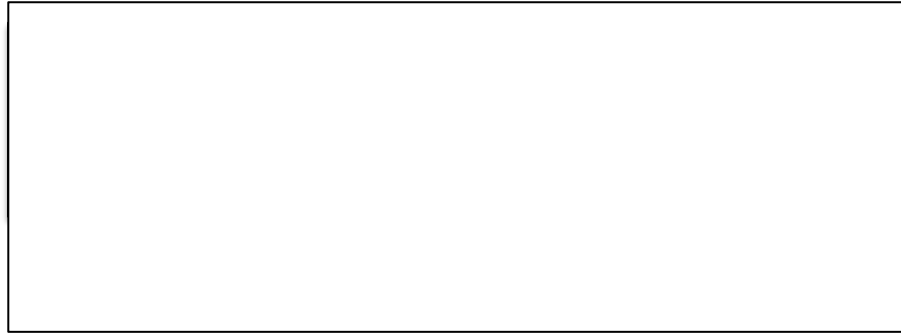
Add PPTase 2
with a red label



AFOSR

Ongoing: create a system with orthogonal reactivity

Adding AcpH erases the slide



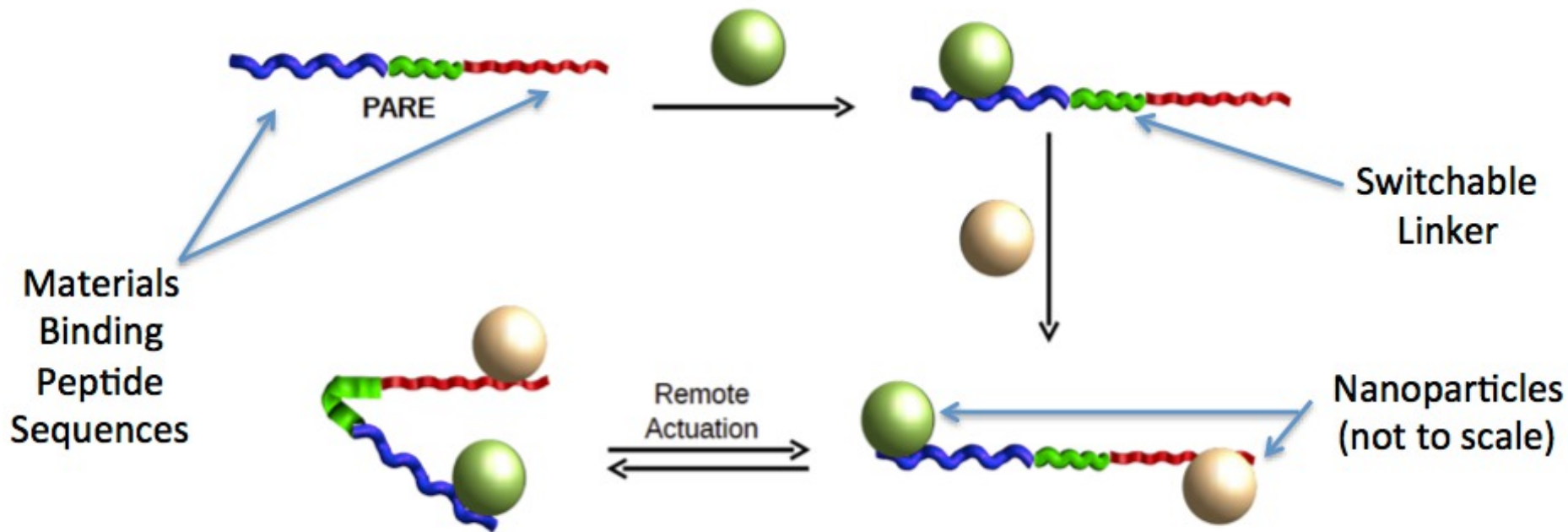
We are applying related ideas to problem 2

1. Finding minimized substrates for a pair of protein-modifying enzymes.

– Joint work with Nathan Gianneschi, Michael Burkart, Michael Gilson

2. Finding specific binders for a given pair of target materials

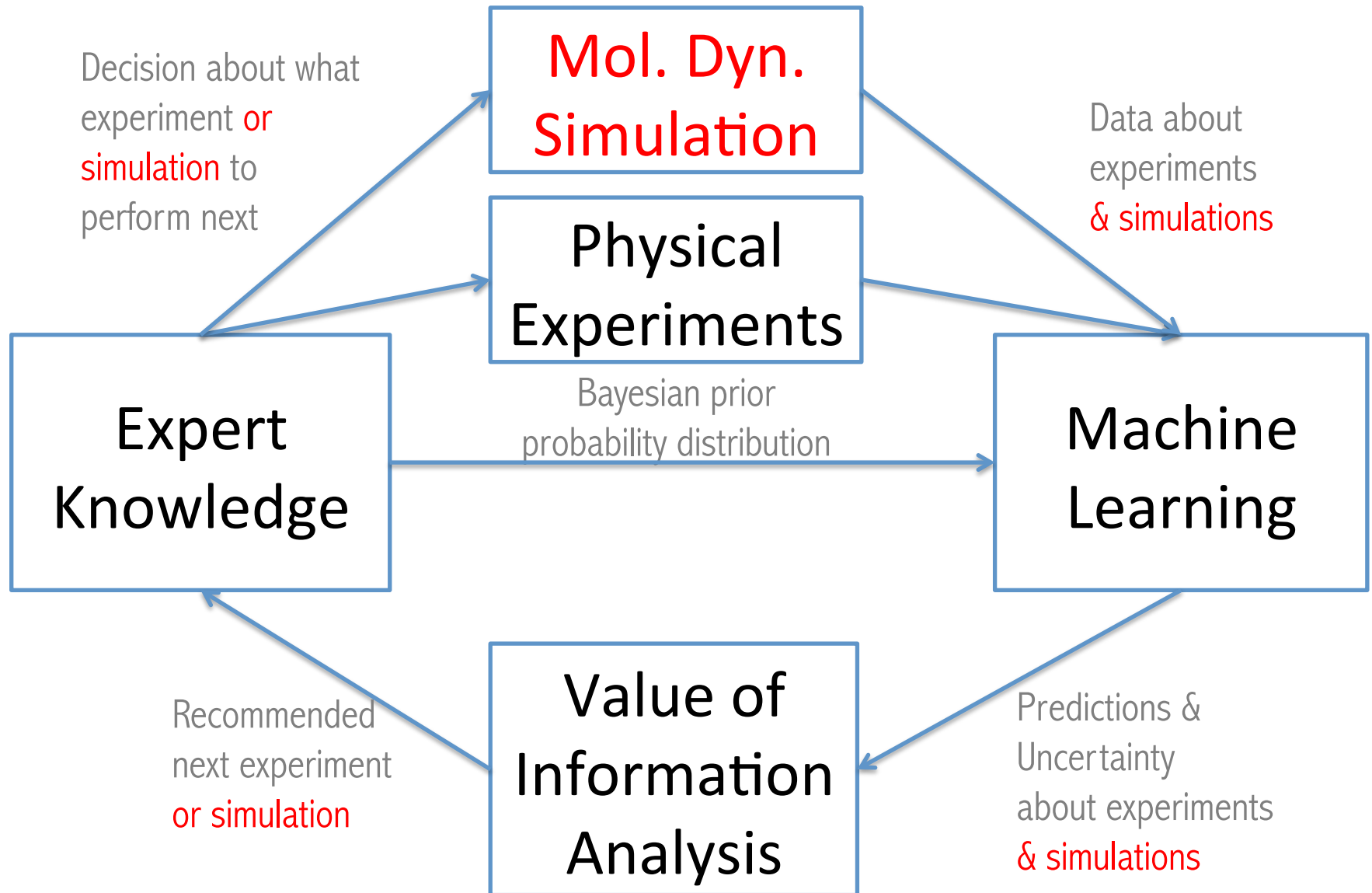
– Joint work with Tiff Walsh and Marc Knecht, working with Paras Prasad, Mark Swihart, and Aidong Zhang.



2. Finding specific binders for a given pair of target materials

- Joint work with Tiff Walsh and Marc Knecht, working with Paras Prasad, Mark Swihart, and Aidong Zhang.

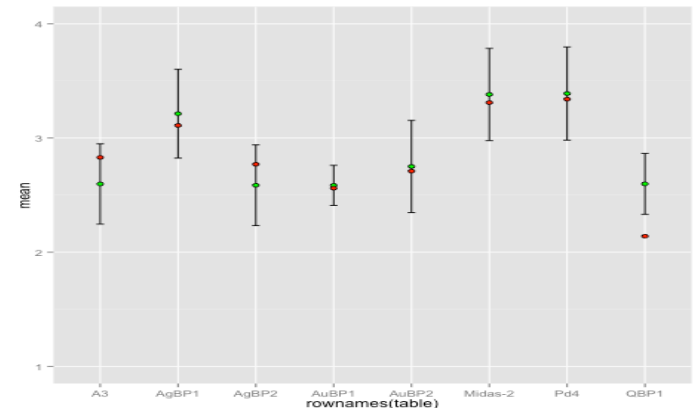
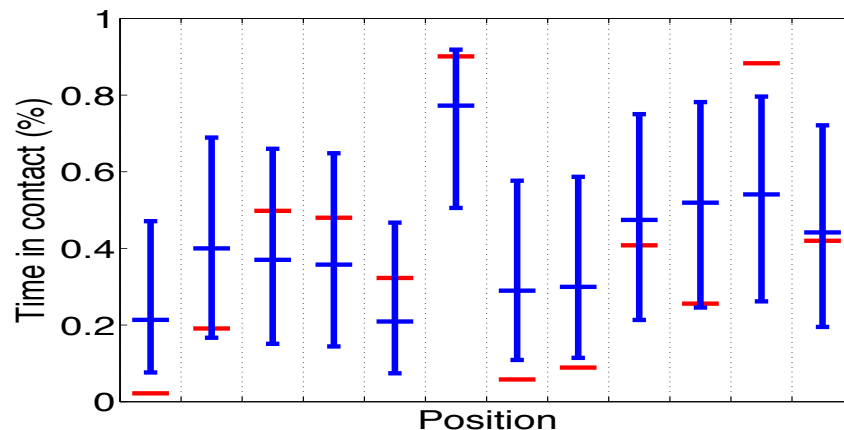
In problem 2, we can also run a molecular dynamics simulation



We are extending these ideas to include simulations as another source of information

We have built fast machine learning methods for predicting:

- The output of the MD simulations: contact residues; configurational entropy.



- The output of the experiment: free energy of binding.
- We are now testing & improving these methods.

We are extending these ideas to include simulations as another source of information

- We have designed value of information methods for recommending:
 1. Whether to perform an MD simulation, or a physical experiment next.
 2. On which peptide to run the simulation/experiment
- We are currently implementing & testing these methods.

Conclusion

- We are developing optimal learning methods for solving two problems in peptide design:
 - Finding minimal peptide substrates
 - Finding specific binders
- These methods:
 1. Reduce the experimental effort required to achieve a desired goal.
 2. Increase the chance of achieving a goal within a given experimental budget.