Optimal Learning for Efficient Sequential Experimental Design in Nano-Bio Research

Natural Materials, Systems and Extremophiles Program Review

January 11, 2013

Warren Powell (PI)
Princeton University

Peter Frazier (Co-PI)
Cornell University

Optimal Learning for Efficient Experimentation in Nanotechnology and Biochemistry

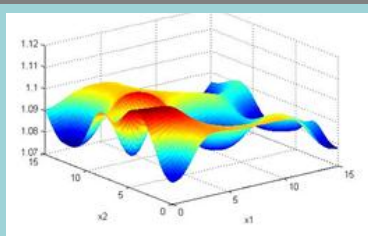
Warren Powell (Princeton), Peter Frazier (Cornell)

STATUS QUO

Current experimental strategy:

- *Most experimentation is guided by experts trying the most promising designs.*
- *Classical design of experiments is batch and ignores domain knowledge.*
- *We need better tools for belief extraction in multiple dimensions*

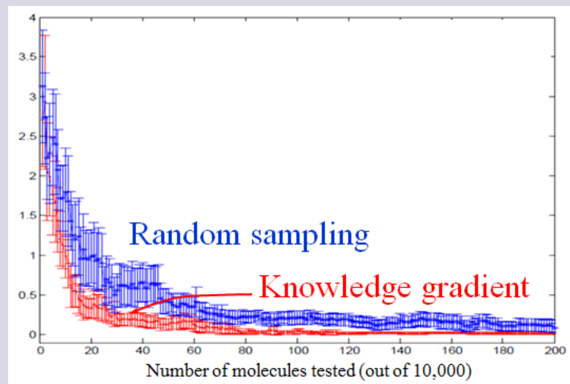
NEW INSIGHTS



- *Knowledge gradient identifies experiments with highest value of information.*
- *Scientists need to understand value of information in an experiment.*

Main achievements

Knowledge gradient significantly increases rate of success in experiments in controlled simulations.



How it works

Computes expected value of an experiment:

$$v_x^{KG,n} = \mathrm{E}^n \max_{x'} \theta_{x'}^{n+1} - \max_{x'} \theta_{x'}^n$$

Challenge is finding expectation of max (best design), and extracting a prior distribution of belief which captures covariances in a transformed parameter space.

QUANTITATIVE IMPACT

- *Knowledge gradient dramatically accelerated the identification of the correct belief model.*

Planned next steps

- *Interactive belief extraction model should capture priors in transformed parameter space.*
- *Compute knowledge gradient for wider range of belief models.*

END-OF-PHASE GOAL

- *Provide interactive environment for belief extraction*
- *Show regions which contribute highest value of information*
- *Simulate experiments to estimate likelihood of success.*

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) We have begun working with scientists to develop a better understanding of the challenges they face. This is a necessary prerequisite before communicating key ideas to this community.

Goal 2) We are working with the Prasad team based at Buffalo, the Wagner/Sturm group at Princeton, and Clancy group at Cornell. We have also initiated discussions with other groups and teams at AFRL and Northwestern.

Goal 3) Progress includes:

- We have begun the development of a web-based belief extraction tool which allows a user to create a family of functions which represent possible behaviors.
- We have begun working with several teams of researchers at Princeton, Cornell, and to understand the research challenges.
- A new informatics strategy for finding peptides with specific binding properties is being developed for use by Paras Prasad's team.

Goal 4) An integrated web-based tool is envisioned for belief extraction, experimental guidance and simulation-based testing. This is being designed, and will build on the web-based extraction tool under development.

Transitions

- None to date.

Interactions with other groups

- We are working with Paras Prasad (Buffalo) and his multi-university team funded by AFOSR 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 interacting with Rajesh Naik’s group at AFRL to obtain data in support of the work with Paras Prasad, and for possible future work on other peptide design problems.
- We are working with the Wagner/Sturm group at Princeton University on amorphous silicon thin-film transistors for organic LEDs to guidance very time-consuming experiments.
- We are interacting with Paulette Clancy’s group at Cornell to show her & her group how to apply optimal learning to a variety of design problems in semiconductor materials.
- We have had preliminary discussions, with the following individuals and groups, about the use of optimal learning in their projects:
  - Chad Mirkin’s group at Northwestern
  - Rajesh Naik’s group at AFRL
  - Rollie Dutton at AFRL

Publications

- Under review/nearing completion:
  - Arta Jamshidi and W. B. Powell, “A Recursive Semi-parametric Approximation Method using Dirichlet Clouds and Radial Basis Functions,” (in preparation, nearing completion).
  - E. Barut and W. B. Powell, “Optimal Learning for Sequential Sampling with Non-Parametric Beliefs,” under final review J. Global Optimization.
  - Boris Defourny, Ilya O. Ryzhov, W. B. Powell, “Optimal Information Blending with Measurements in the L2 Sphere,” submitted to Mathematics of Operations Research, October 12, 2012.
  - Samuel J. Gershman, P.I. Frazier, David M. Blei, “Distance Dependent Infinite Latent Feature Models,” in review IEEE Trans. Pattern Analysis and Machine Intelligence.
  - Scott C. Clark, Rob Egan, P.I. Frazier, and Zhong Wang, “ALE: a Generic Assembly Likelihood Evaluation Framework for Assessing the Accuracy of Genome and Metagenome Assemblies,” in review at Bioinformatics.
  - Jialei Wang, Scott Clark, P.I. Frazier, “Parallel Global Optimization Using Multipoints Expected Improvement” in preparation, nearing completion.
  - “Optimal Learning for Peptide Design” in preparation with the Prasad team.

Publications

- Appeared in 2012:
  - S.E. Chick and P.I. Frazier, “Sequential Sampling for Selection with Economics of Selection Procedures,” Management Science, vol 58, no 3, pp 550–569, 2012.
  - I.O. Ryzhov, W.B. Powell, and P.I. Frazier “The Knowledge-Gradient Algorithm for a General Class of Online Learning Problems,” Operations Research, vol. 60, pp 180–195, 2012.
  - R. Waeber, P.I. Frazier and S.G. Henderson, “A Framework for Selecting a Selection Procedure,” ACM Transactions on Modeling and Computer Simulation, vol. 22, no. 3, 2012.
  - B. Jedynak, P.I. Frazier, and R. Sznitman, “Twenty Questions with Noise: Bayes Optimal Policies for Entropy Loss,” Journal of Applied Probability, vol 49, pp 114-136, 2012.
  - 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, in press, available online 13 Sep 2012.
  - S. Zhang, P. Hanagal, P.I. Frazier, A.J. Meltzer, D.B. Schneider, “Optimal Patient-specific Post-operative Surveillance for Vascular Surgery,” 7th INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2012), 2012.
  - P.I. Frazier, “Tutorial: Optimization via Simulation with Bayesian Statistics and Dynamic Programming,” Winter Simulation Conference, 2012.
  - P.I. Frazier, Bruno Jedynak, and Li Chen, “Sequential Screening: A Bayesian Dynamic Programming Analysis of Optimal Group-Splitting,” Winter Simulation Conference, 2012.
  - J. Xie, P. I. Frazier, S. Sankaran, A. Marsden, and S. Elmohamed, “Optimization of Computationally Expensive Simulations with Gaussian Processes and Parameter Uncertainty: Application to Cardio-vascular Surgery,” 50th Annual Allerton Conference on Communication, Control, and Computing, 2012.

Transformational/evolutionary

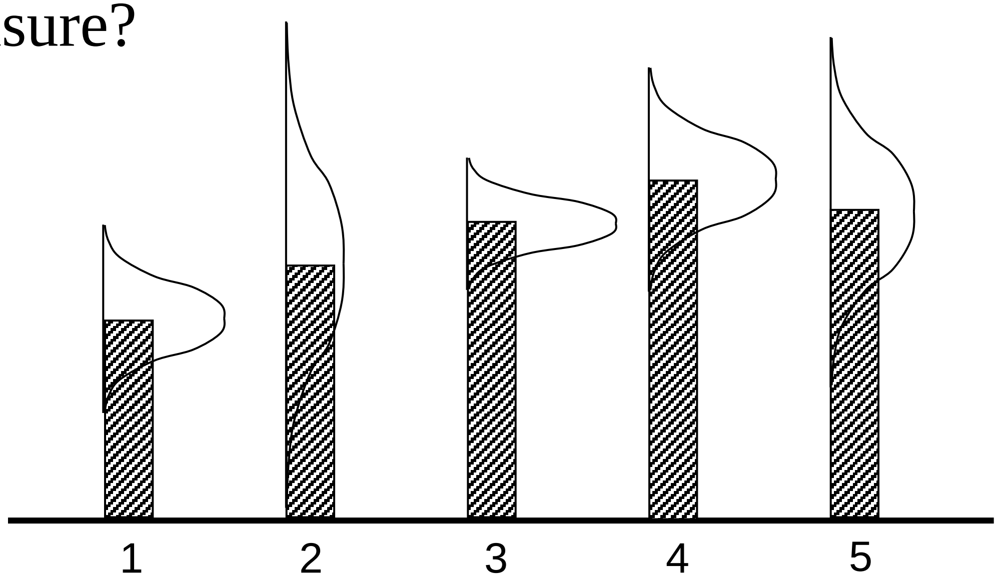
- Our research has the potential for transforming the fundamental style in which research is conducted in the physical sciences, providing a principled strategy to the complex but often ad hoc process of experimental design.
- Our work will, at a minimum, help scientists understand the value of information, which is a distinctly different goal than running experiments with the hope of success.
- It will help scientists work with their beliefs to estimate the likelihood of success from experiments where there is a high level of uncertainty, helping them assess risks and rewards from different experimental strategies.

Outline

- Introduction to optimal learning using the knowledge gradient
- Belief extraction
- Collaboration with the Prasad team

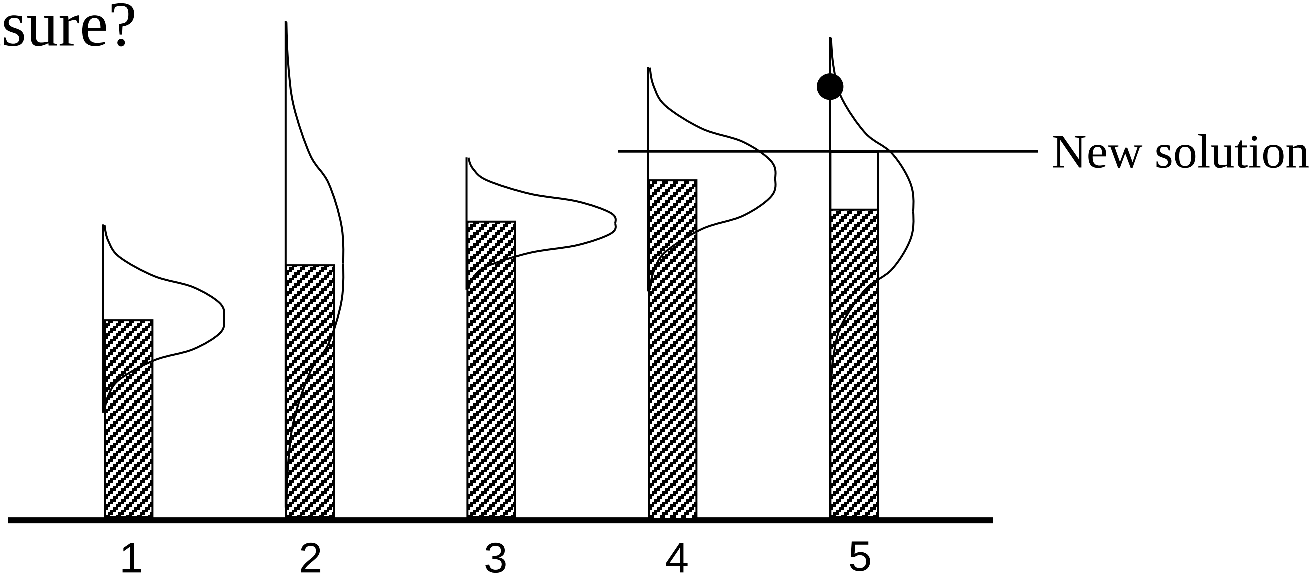
The learning problem

- Assume we have five choices of experiments, with uncertainty in our belief about how well each one will perform.
- If you can perform one experiment, which would you measure?



The learning problem

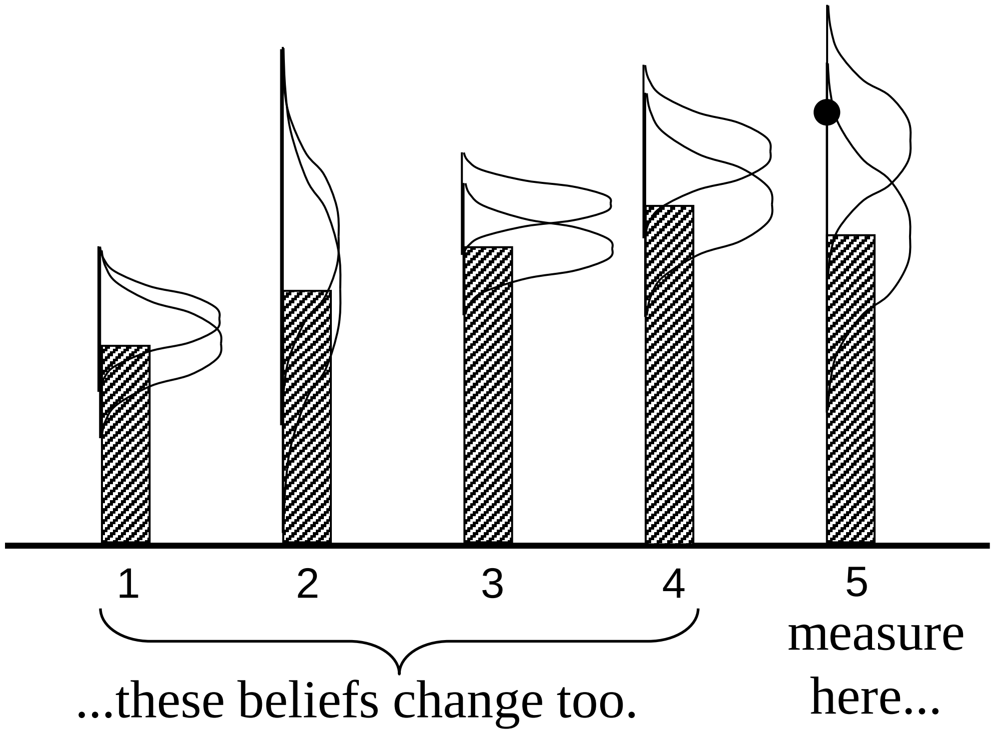
- Assume we have five choices of experiments, with uncertainty in our belief about how well each one will perform.
- If you can perform one experiment, which would you measure?



- The value of information is the expected improvement in a design as a result of an experiment, which requires striking a balance between potential performance and uncertainty.

Correlated beliefs

- An important problem class involves *correlated beliefs* – measuring one alternative tells us something other alternatives.



- Correlated beliefs allow us to dramatically reduce the number of experiments that need to be run.

Learning with correlated beliefs

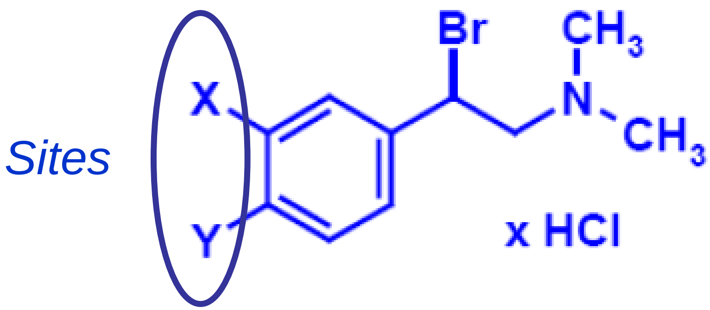
- Similarities in molecular structure implies that groups of molecules will behave similarly
  - Each row/column represents a type of molecule.
  - Chemist used judgment to estimate the correlation coefficient for groups of molecules with similar structure

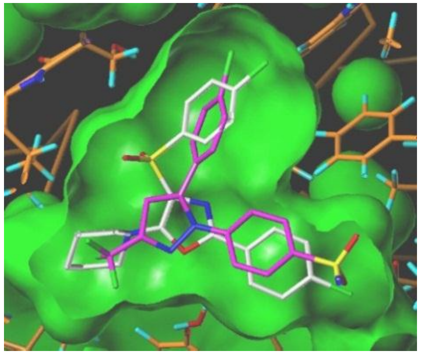
Index	PT-1-11	PT-1-12	PT-1-13	PT-1-14	PT-1-15	PT-1-16	PT-1-17	PT-1-18	PT-1-19	PT-1-20	PT-1-21	PT-1-22	PT-1-23
PT-1-11	0.27878	0.048776	0.014781	0.10878	0.20878	0.20878	0.088776	0.048776	0.014781	0.048776	0.014781	0.014781	0.014781
PT-1-12	0.048776	0.27878	0.014781	0.048776	0.048776	0.048776	0.048776	0.048776	0.014781	0.27878	0.014781	0.014781	0.24478
PT-1-13	0.014781	0.014781	0.53418	0.014781	0.014781	0.014781	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
PT-1-14	0.10878	0.048776	0.014781	0.27878	0.10878	0.10878	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
PT-1-15	0.20878	0.048776	0.014781	0.10878	0.27878	0.22878	0.068776	0.048776	0.014781	0.048776	0.014781	0.014781	0.014781
PT-1-16	0.20878	0.048776	0.014781	0.10878	0.22878	0.27878	0.068776	0.048776	0.014781	0.048776	0.014781	0.014781	0.014781
PT-1-17	0.088776	0.048776	0.014781	0.13878	0.068776	0.068776	0.27878	0.048776	0.014781	0.048776	0.014781	0.014781	0.014781
PT-1-18	0.048776	0.048776	0.24478	0.048776	0.048776	0.048776	0.048776	0.27878	0.014781	0.048776	0.24478	0.014781	0.014781
PT-1-19	0.014781	0.014781	0.074176	0.014781	0.014781	0.014781	0.014781	0.014781	0.30418	0.014781	0.074176	0.30418	0.074176
PT-1-20	0.048776	0.27878	0.24478	0.048776	0.048776	0.048776	0.048776	0.048776	0.014781	0.50878	0.014781	0.014781	0.47478
PT-1-21	0.014781	0.014781	0.30418	0.014781	0.014781	0.014781	0.014781	0.24478	0.074176	0.014781	0.30418	0.074176	0.074176
PT-1-22	0.014781	0.014781	0.074176	0.014781	0.014781	0.014781	0.014781	0.014781	0.30418	0.014781	0.074176	0.30418	0.074176
PT-1-23	0.014781	0.24478	0.30418	0.014781	0.014781	0.014781	0.014781	0.014781	0.074176	0.47478	0.074176	0.074176	0.53418
PT-1-33	0.01478	0.01478	0.074176	0.01478	0.01478	0.01478	0.01478	0.01478	0.074176	0.01478	0.074176	0.074176	0.074176
PT-1-38	0.098776	0.048776	0.014781	0.13878	0.068776	0.068776	0.27878	0.048776	0.014781	0.048776	0.014781	0.014781	0.014781
PT-1-39	0.01478	0.01478	0.074176	0.01478	0.01478	0.01478	0.01478	0.01478	0.31418	0.01478	0.074176	0.31418	0.074176
PT-1-41	0.01478	0.24478	0.31418	0.01478	0.01478	0.01478	0.01478	0.01478	0.074176	0.48478	0.074176	0.074176	0.54418

Highest covariance along the diagonal.

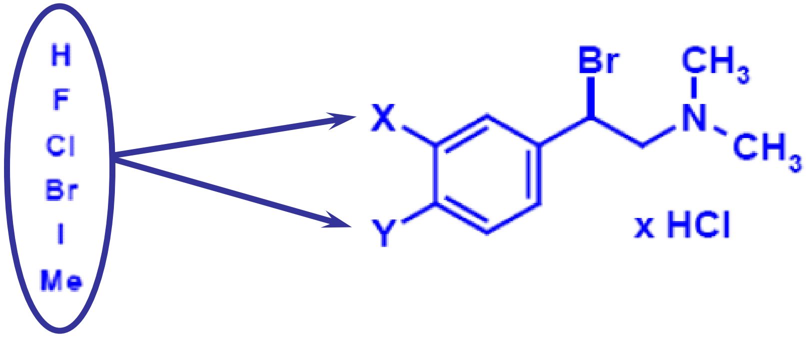
Optimal learning in materials science

- Molecular design
  - Designing molecules is time consuming and expensive





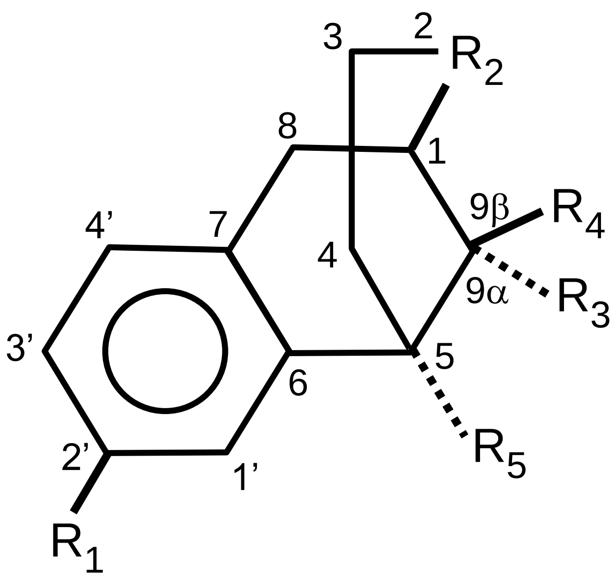
  - X and Y are *sites* where we can hang *substituents* to change the behavior of the molecule



  - The goal is to find the best set of substituents.

Optimal learning in materials science

- A more complex molecule:



Potential substituents:

F	OCH_3
OH	CH
CH_3	NO
$OCOCH_3$	CI
OCOCH	OCOCH

- From this base molecule, we created problems with 10,000 compounds, and one with 87,120 compounds.
- We need to find the best with a minimum number of experiments.

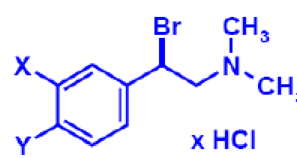
Optimal learning in materials science

- We express our belief using a linear, additive QSAR model
  - $X_{ij}^{m} = 1$ if substituent j is at site i, 0 otherwise.
  - $Y = \theta_0 + \sum_{sites\ i} \sum_{substituents\ j} \theta_{ij} X_{ij}$ QSAR belief model

Hugo Kubinyi, www.kubinyi.de

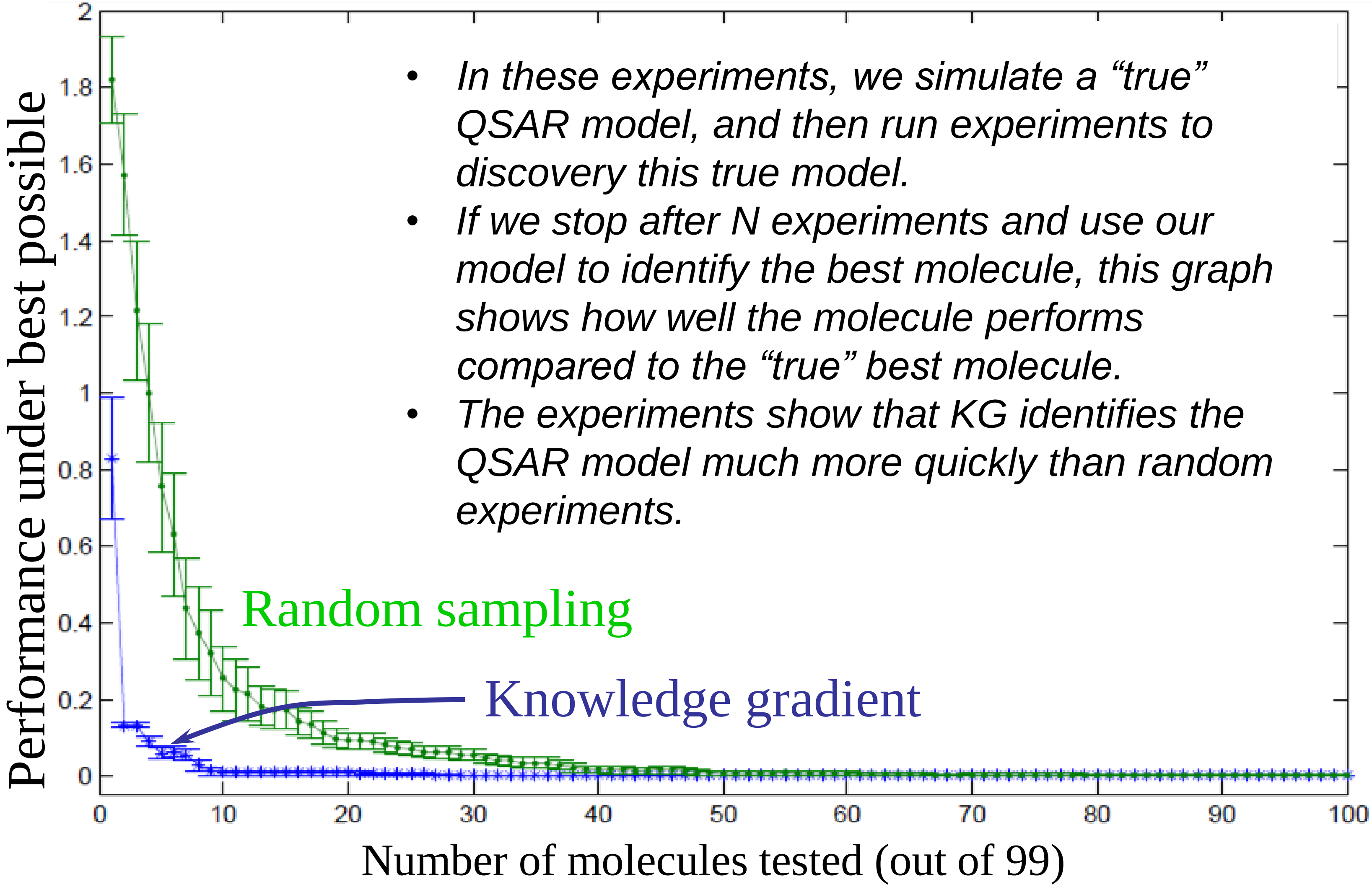
meta (X)	para (Y)	log 1/C obs.	meta- F	Cl	Br	I	Me	para- F	Cl	Br	I	Me	log 1/C calc.
H	H	7.46											7.82
H	F	8.16						1					8.16
H	Cl	8.68							1				8.59
H	Br	8.89								1			8.84
H	I	9.25									1		9.25
H	Me	9.30										1	9.08
F	H	7.52	1										7.52
Cl	H	8.16		1									8.03
Br	H	8.30			1								8.26
I	H	8.40				1							8.40
Me	H	8.46					1						8.28
Cl	F	8.19		1				1					8.37
Br	F	8.57			1			1					8.60
Me	F	8.82					1	1					8.62
Cl	Cl	8.89		1					1				8.80
Br	Cl	8.92			1				1				9.02
Me	Cl	8.96					1		1				9.04
Cl	Br	9.00		1						1			9.05
Br	Br	9.35			1					1			9.28
Me	Br	9.22					1			1			9.30
Me	Me	9.30					1					1	9.53
Br	Me	9.52			1							1	9.51

Matrix for Free Wilson Analysis



Optimal learning in materials science

- Knowledge gradient versus random sampling for 99 compounds

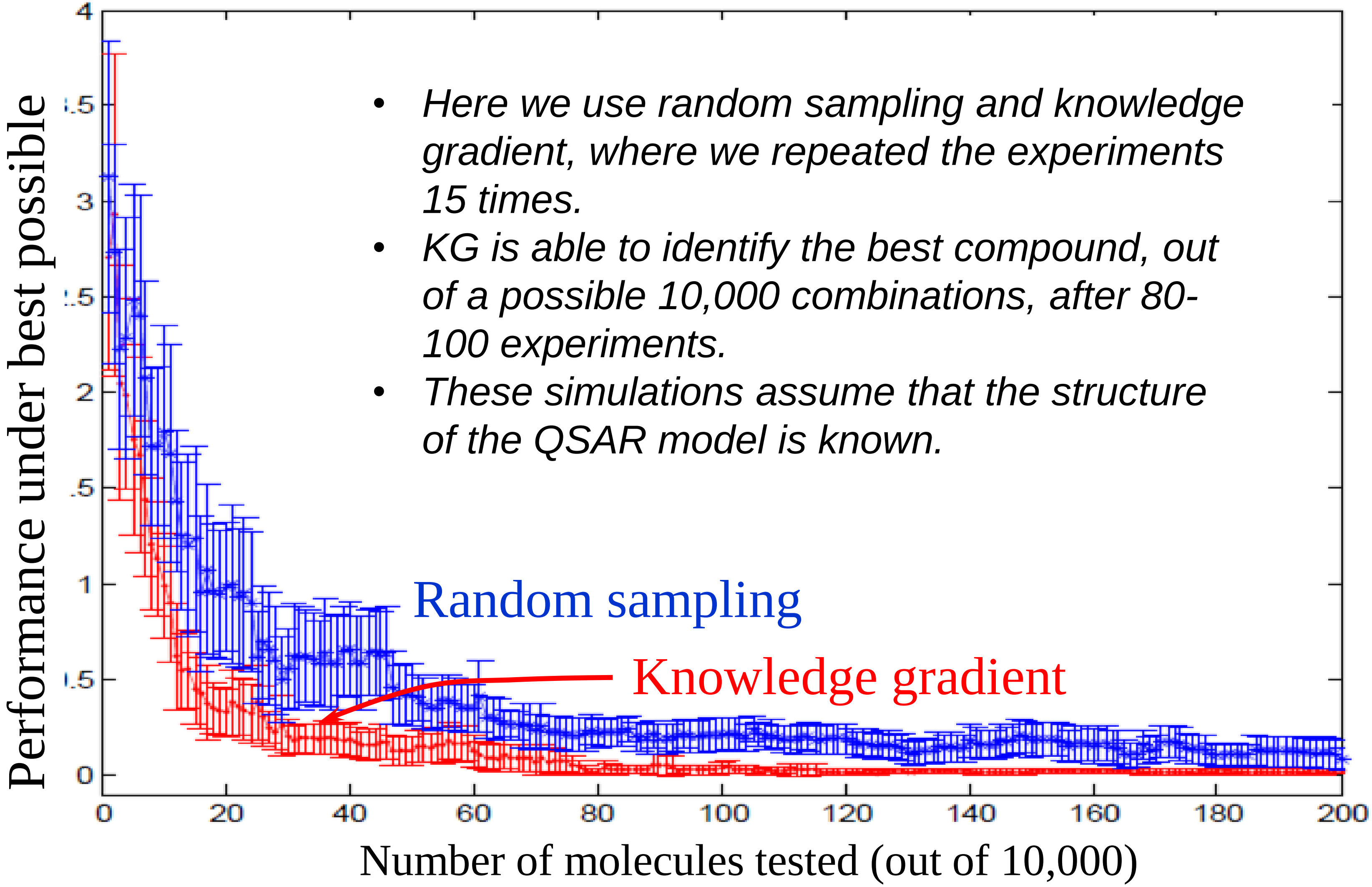


- *In these experiments, we simulate a "true" QSAR model, and then run experiments to discovery this true model.*
- *If we stop after N experiments and use our model to identify the best molecule, this graph shows how well the molecule performs compared to the "true" best molecule.*
- *The experiments show that KG identifies the QSAR model much more quickly than random experiments.*

Optimal learning in materials science

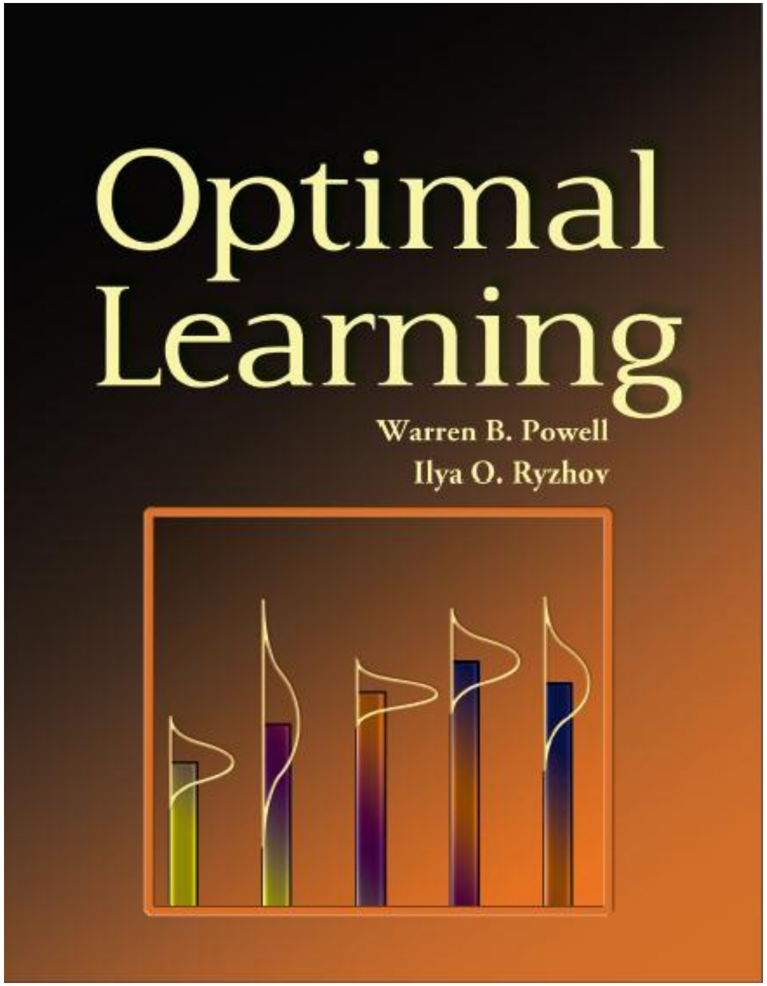
- Knowledge gradient vs random sampling for 10,000 compounds

- *Here we use random sampling and knowledge gradient, where we repeated the experiments 15 times.*
- *KG is able to identify the best compound, out of a possible 10,000 combinations, after 80-100 experiments.*
- *These simulations assume that the structure of the QSAR model is known.*



New book!

- New book on *Optimal Learning*
  - Published by John Wiley in 2012.
  - First 12 chapters are at an advanced undergraduate level.
  - Funded by AFOSR
- Synthesizes communities:
  - Ranking and selection
  - Multiarmed bandits
  - Stochastic search
  - Simulation optimization
  - Global optimization
  - Experimental design



http://optimallearning.princeton.edu/

Optimal learning

- Research challenges for the nano-bio community
  - Belief extraction
    - The scientific research community possesses a tremendous amount of knowledge about the behavior of materials
    - We need to capture this domain knowledge, including both *what you know* and *how well you know it.*
  - Guidance
    - We need to present evidence to convince you to run experiments.
    - This guidance needs to be interactive, working with your domain knowledge.
  - Nano-bio research poses new challenges that have not yet been addressed by the optimal learning community.
    - Wide array of belief models that we have not considered.
    - Challenges such as dealing with massive arrays

Outline

- Introduction to optimal learning using the knowledge gradient
- Belief extraction
- Collaboration with the Prasad team

Belief extraction

- The challenge
  - Learning what these domain experts know
  - We have to extract *what they know* and *how well they know it.*
  - The goal is to guide their research, and help them understand the potential and risks of different research strategies.

Belief extraction

- Challenge: Nanocrystalline silicon for high-performance/low-power transistor circuit technology on flexible substrates
  - Profs. Sigurd Wagner, Jim Sturm, with Ph.D. student Ting Liu
  - The silicon nitride and amorphous silicon are grown in a standard plasma-enhanced chemical vapor deposition (PECVD) system subject to choices of a number of parameters.

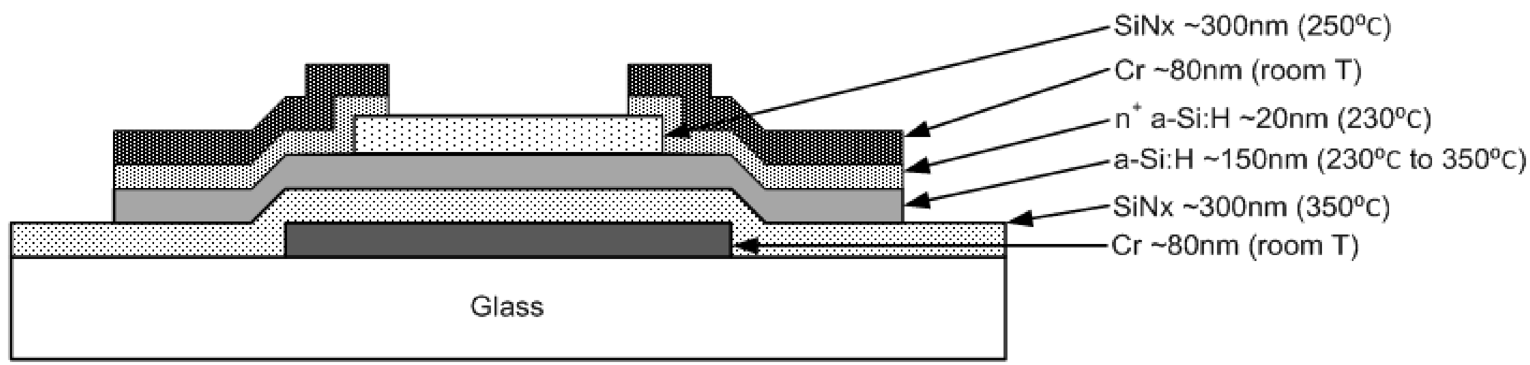


Fig. 4 Schematic cross-section of back-channel passivated TFT structure

  - Considerable variation in the results from one run to the next due to unknown causes.

Belief extraction

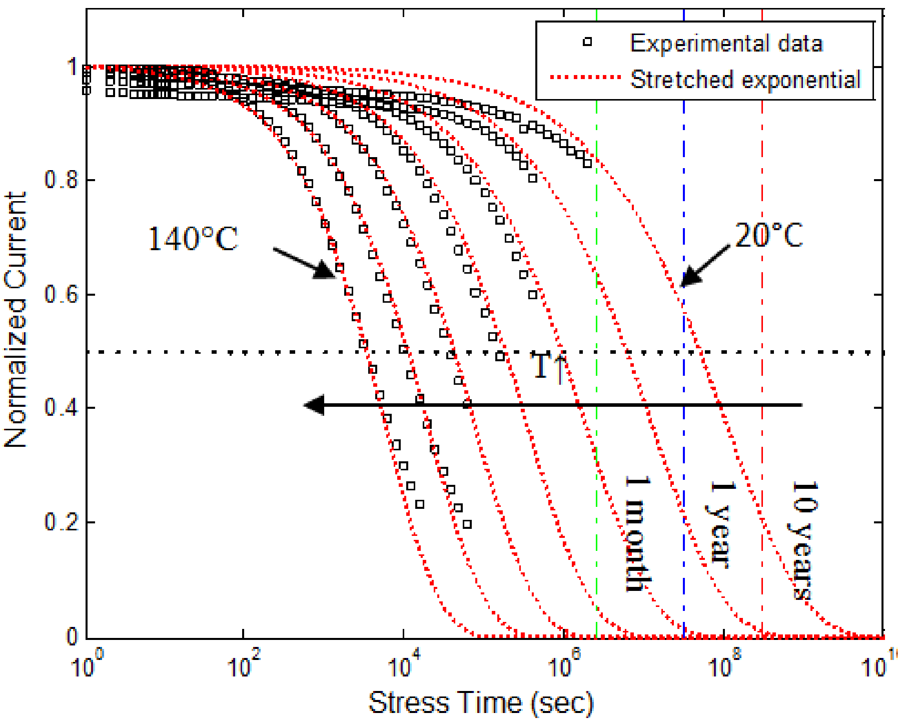
❑ Controllable parameters

- Flow rates of the two source gases [sccm]
- Pressure during nanocrystalline silicon film growth [in Pascal]
- Radiofrequency power that is fed into the reactor [in Watts]
- Flow rates of the two source gases [in volume per time]
- Pressure during insulator film growth [in Pascal]
- Radiofrequency power that is fed into the reactor [in Watts]
- Temperature of the substrate on which the insulator is grown [in °C]

	Recipes		Stage I	
	SiH4/NH3/H2	SiH4/H2	ΔV_{T1} (V)	t_0 (s)
C7	5/50/220, 2W	20/100, 2W, 350C	0.08	1.1
C5	5/50/220, 2W	20/40, 2W, 350C	0.2	0.4
T50	5/50/220, 2W	20/200, 2W, 320C	0.08	30.8
T13	6/60/300, 5W	16/200, 4W, 250C	0.08	10.3
T12	6/90/300, 5W	16/200, 4W, 250C	0.15	1.6
T14	6/120/300, 5W	16/200, 4W, 250C	0.14	1.3
Y3-2	10/150, 5W	16/200, 4W, 250C	0.1	<<1
Y3-3	6/160, 5W	16/200, 4W, 250C	0.1	<<1
Y3-1	10/100, 5W	16/200, 4W, 250C	0.3	<<1

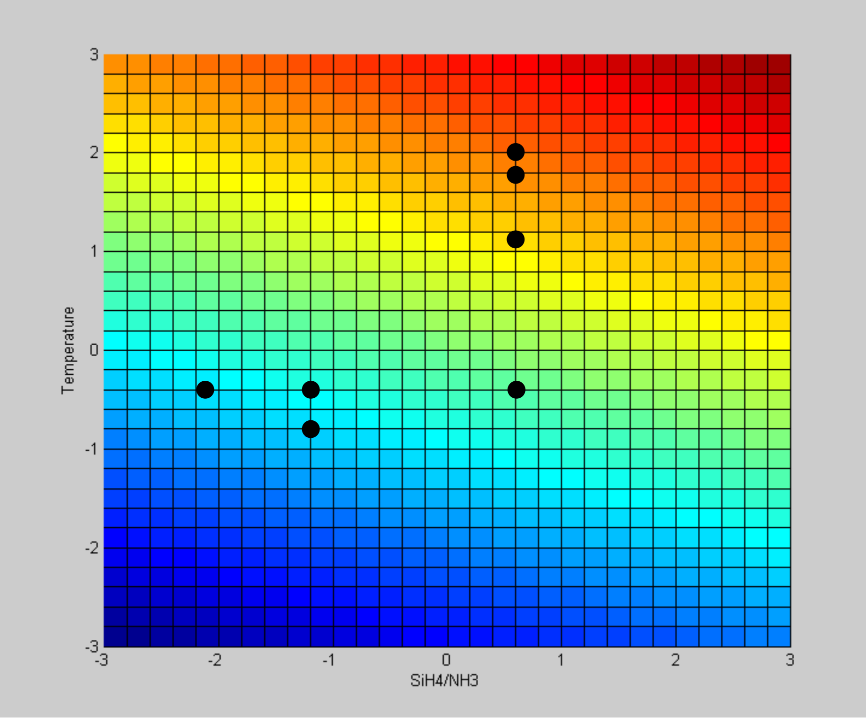
Belief extraction

- Estimating the lifetime
  - Under normal temperatures, material is designed to last several years
  - Not possible to observe such long lifetimes in the lab
  - Deterioration is accelerated under high temperatures.
  - Use observations of deterioration at different temperatures to estimate lifetime under normal temperature.



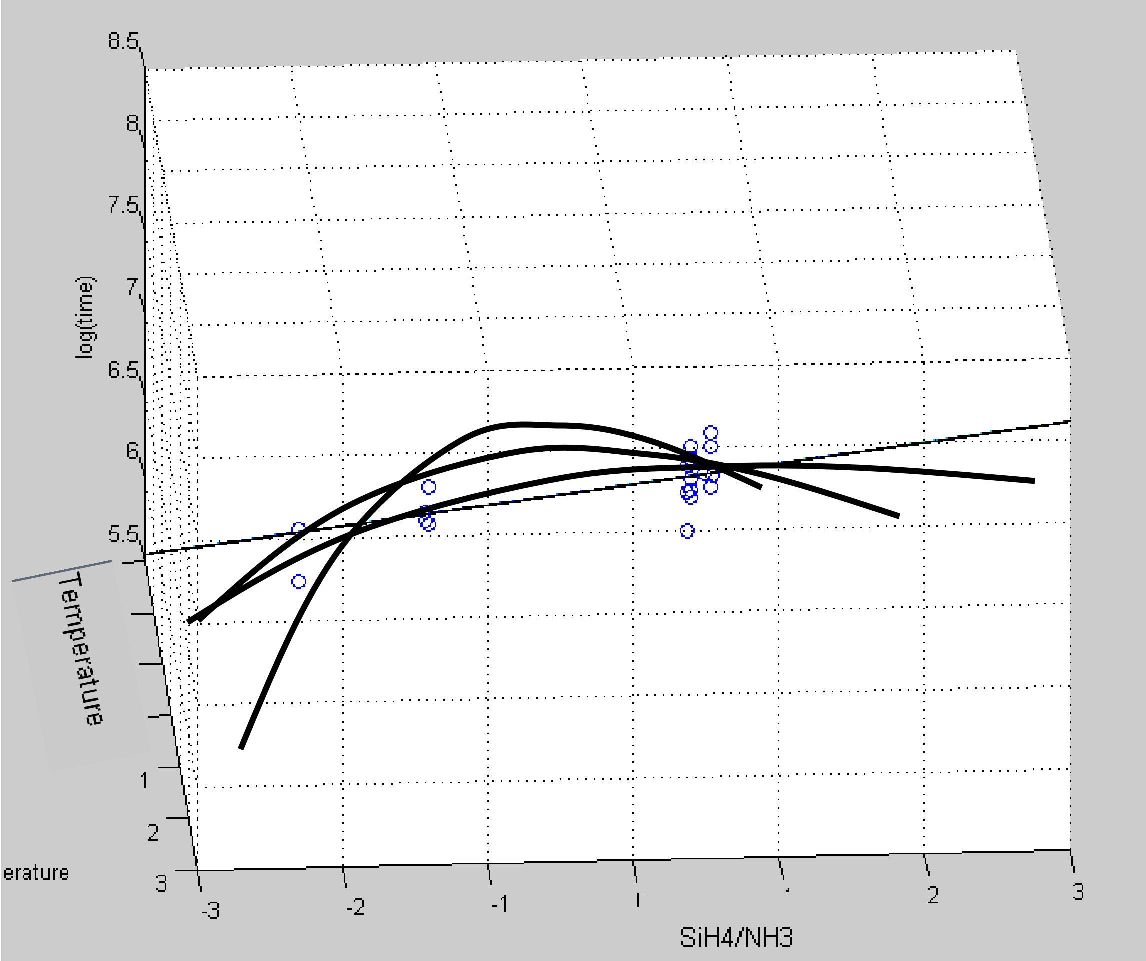
Belief extraction

- Pattern of past experiments
  - Data collected over two years
  - Tendency was to vary one parameter at a time to simplify analysis
  - Statistically, you learn more if you vary all the parameters at once. But how to choose?



Belief extraction

- Data suggests a possible linear relationship
  - But other relationships may be possible



Belief extraction

- ❑ The expert sketches a series of possible models

Function Sketcher for Belief Modeling

Optimal Learning at Princeton

Presets | Function Editor | Import/Export

X-AXIS: Lin | Log
Y-AXIS: Lin | Log
FUNCTION EDITING: Copy | Edit

FUNCTIONS

X	F(X)
0.8645701460551536	1.907901801918402
3.853234222442075	2.5794210581852304
6.898317757009347	0.6799999999999998

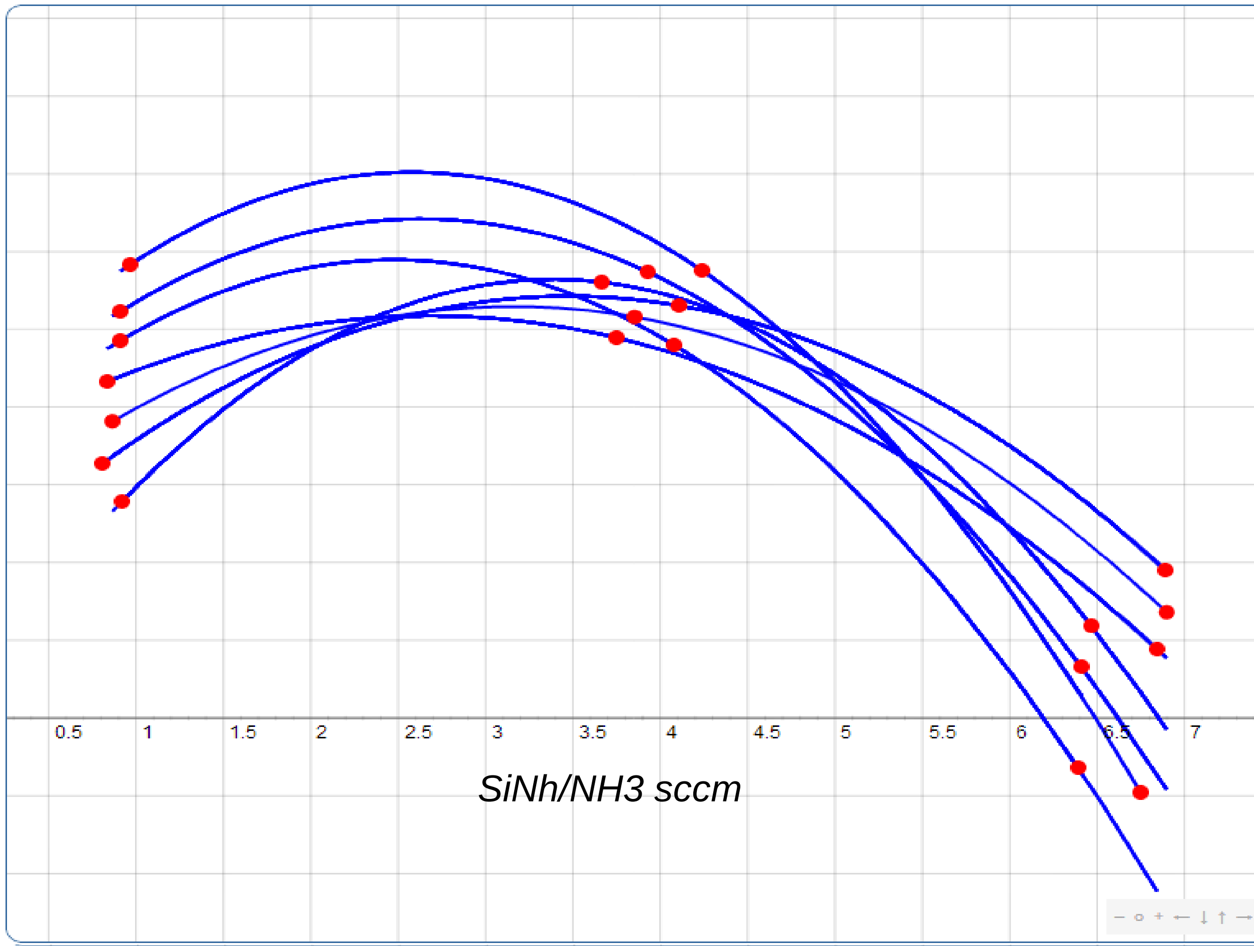
X	F(X)
0.9100191540337695	2.615149978310794
3.9284601666825427	2.869037540122373
6.410916326617983	0.3301619220065823

X	F(X)
0.919422397063828	1.392728384403191
3.6651693618409054	2.8032148389119635
6.467335784798333	0.5934527268482198

X	F(X)
0.8347932097933016	2.1637943128679864
3.7497985491114316	2.4458916037697414
6.843465506000673	0.4430008383672841

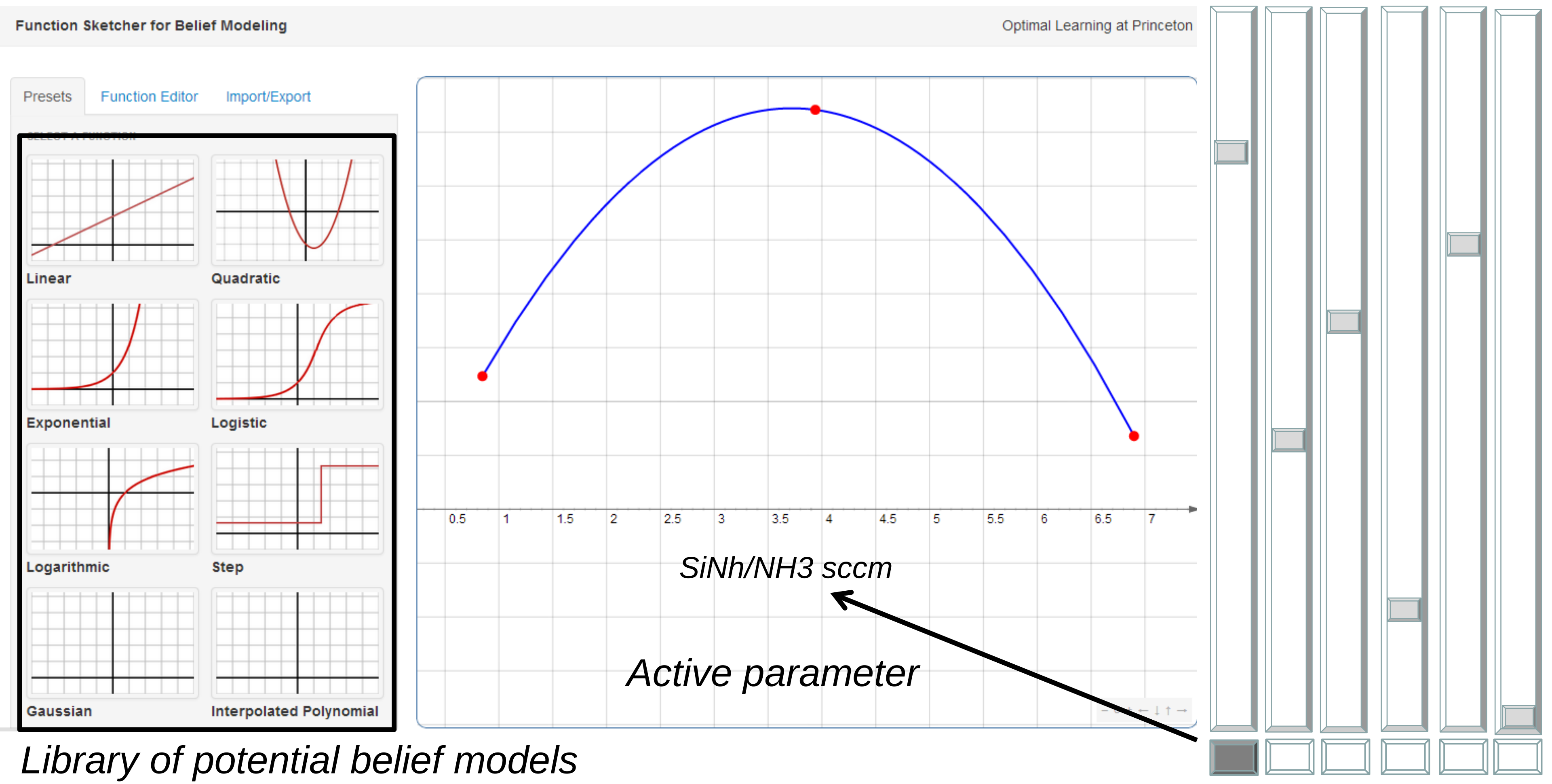
Reset | Get Covariances

Shift-drag to pan. Zoom using mouse wheel.



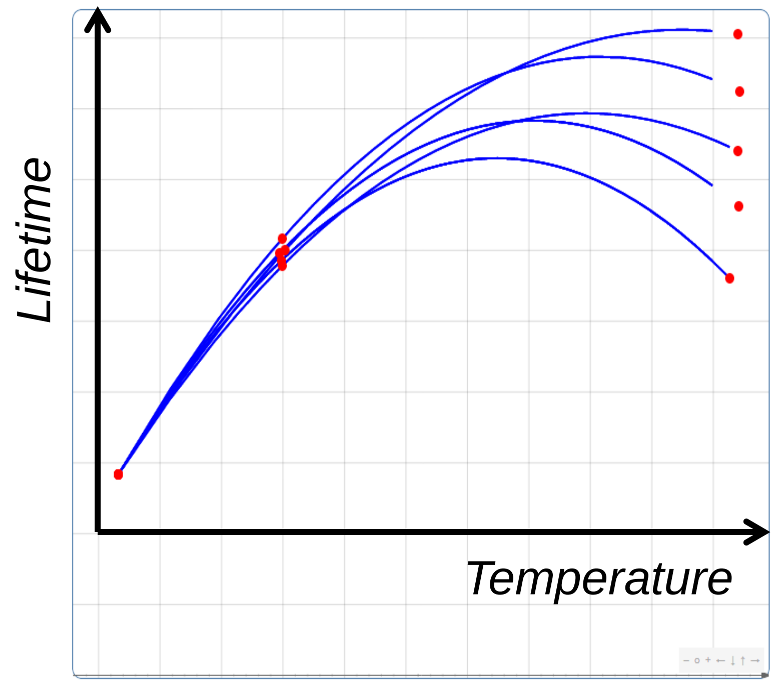
Belief extraction interface

To handle multiple dimensions, we display the response surface for one parameter while fixing the other parameters.
After recording beliefs, you can change the settings of other parameters and repeat, capturing more complex surfaces.
The model will fill in to create a full surface.

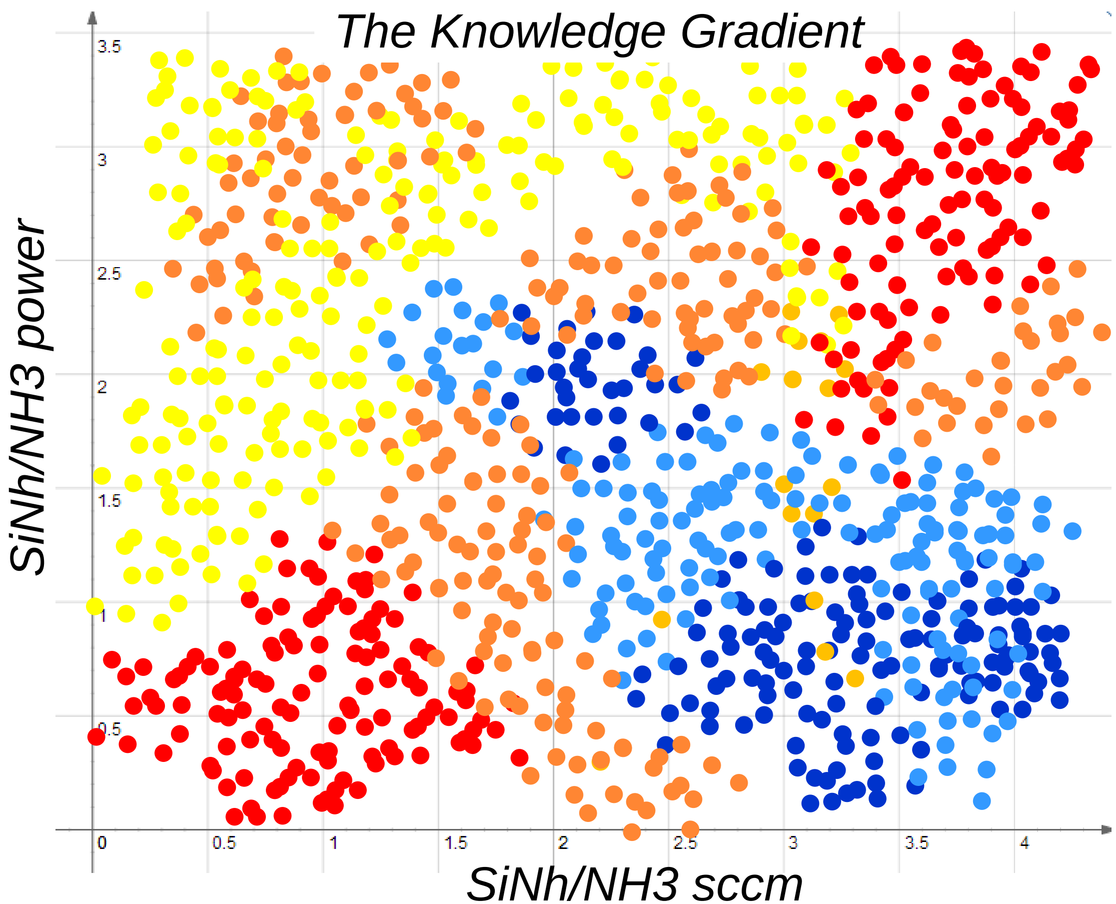


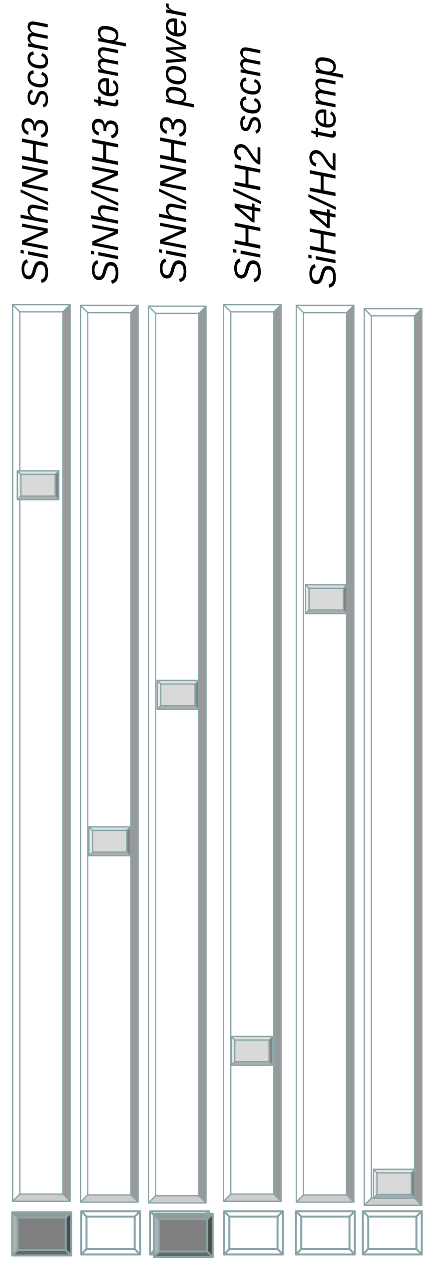
Belief extraction

- Creating a belief curve
  - There are many observations of performance in the region for lower temperatures.
  - Interactive tool allows you to express your uncertainty in the performance over higher temperatures.
  - The knowledge gradient will then identify the experiments that produce the highest value.



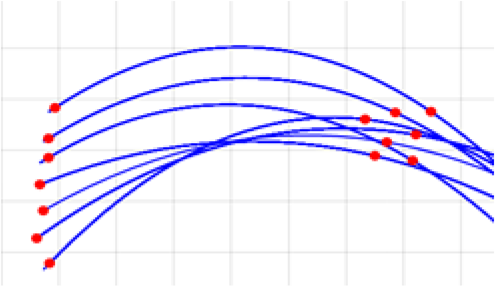
- Identifying the next experiments to run
  - Red points indicate parameter settings that produce the highest value of information
  - The scientist then chooses the most promising experiment to conduct next, balancing the value of information with other considerations.

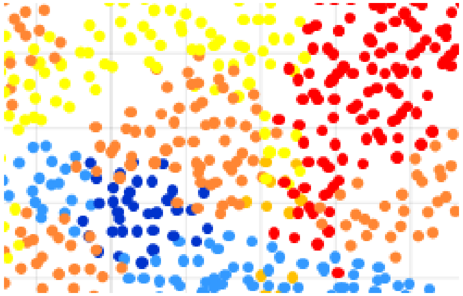


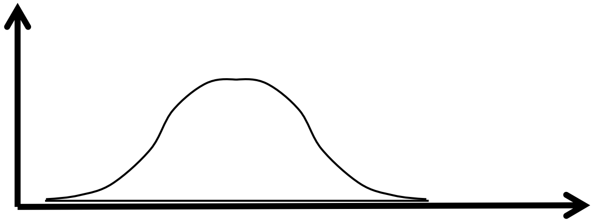


Next steps

- The Optimal Learning Experimental Simulator
  - Belief extraction module captures beliefs about response behavior.
  - Knowledge gradient identifies experiments with the highest value of information.
  - Risk assessment simulator – We simulate a fixed experimental budget hundreds of times to estimate the probability of eventual success.







Outline

- Introduction to optimal learning using the knowledge gradient
- Belief extraction
- Collaboration with the Prasad team

Collaboration with the Prasad Team

- We have begun a collaboration with this team, also funded by AFOSR Natural Materials, Systems and Extremophiles.
  - Paras Prasad (PI, Buffalo)
  - Tiff Walsh (Deakin)
  - Marc Knecht (Miami)
  - Mark Swihart (Buffalo)
  - Aidong Zhang (Buffalo)

- The team's goal is to develop a versatile 3D bio-mediated nanoparticle assembly paradigm for the production of reconfigurable biological nanoassemblies with useful photonic, electronic plasmonic and magnetic properties.
  - The use of peptides as the biological linkers between these materials is central to this goal.

Collaboration with the Prasad Team

- **We are supporting their goals with optimal learning**

  - **Prasad team's immediate goal**: Given two materials, e.g., gold and iron oxide, identify peptides that bind strongly to the first material (e.g., gold), and weakly to the second (e.g., iron oxide).

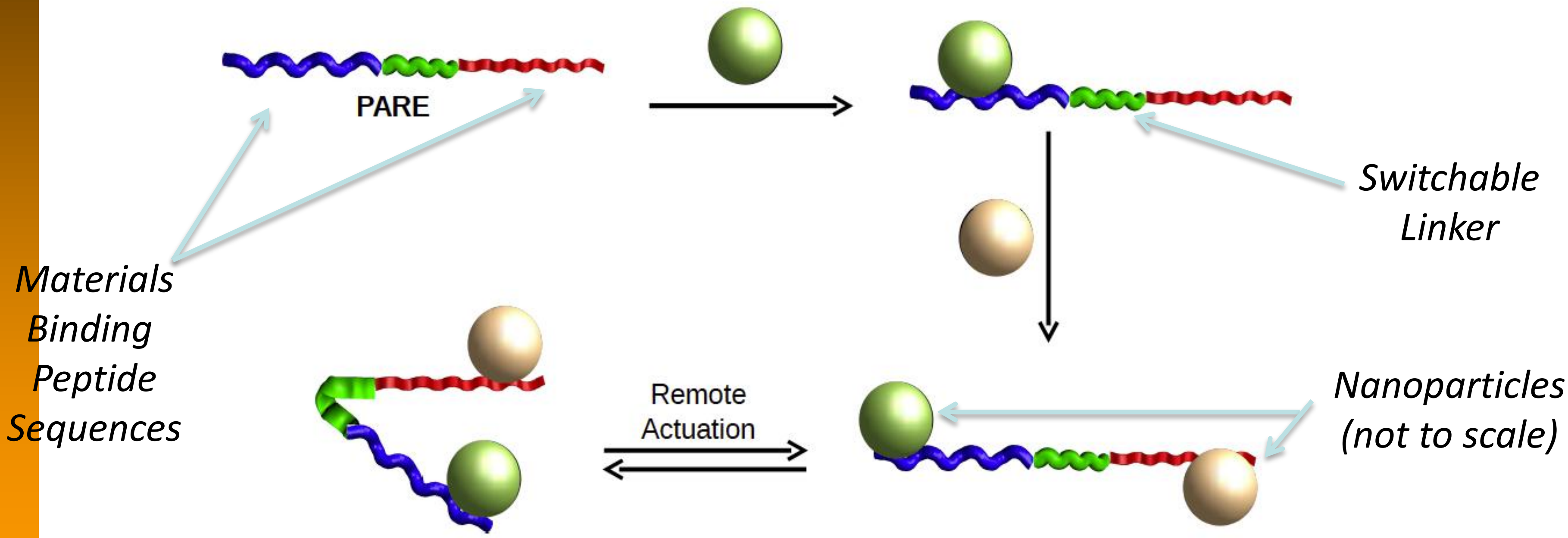
  - **Optimal learning supports this goal:** We will use optimal learning to suggest which peptides would be most informative to synthesize and test in this identification task.

Collaboration with the Prasad Team

- Team's medium-term goal
  - Identifying peptides that bind strongly to the first material, e.g., gold, and weakly to the second, e.g., iron oxide, will support the Prasad team's medium-term goal:
  - A peptide identified in the first step will form one end of a linker molecule (denoted a PARE) that can connect the two materials in a specific manner.
  - The other end of the PARE will be a peptide with opposite binding properties: weak to gold; and strong to iron oxide.
  - A middle piece of the link will allow control of distance between the ends through, e.g., pH or tempurature.
  - Such links will provide great control over how gold and iron oxide are combined at the nano scale.

Prasad team's Overall Strategy for Bio-nanocombinatorics

Will use a library of material-binding peptides connected by switchable linkers to assemble nanoparticles into reconfigurable assemblies

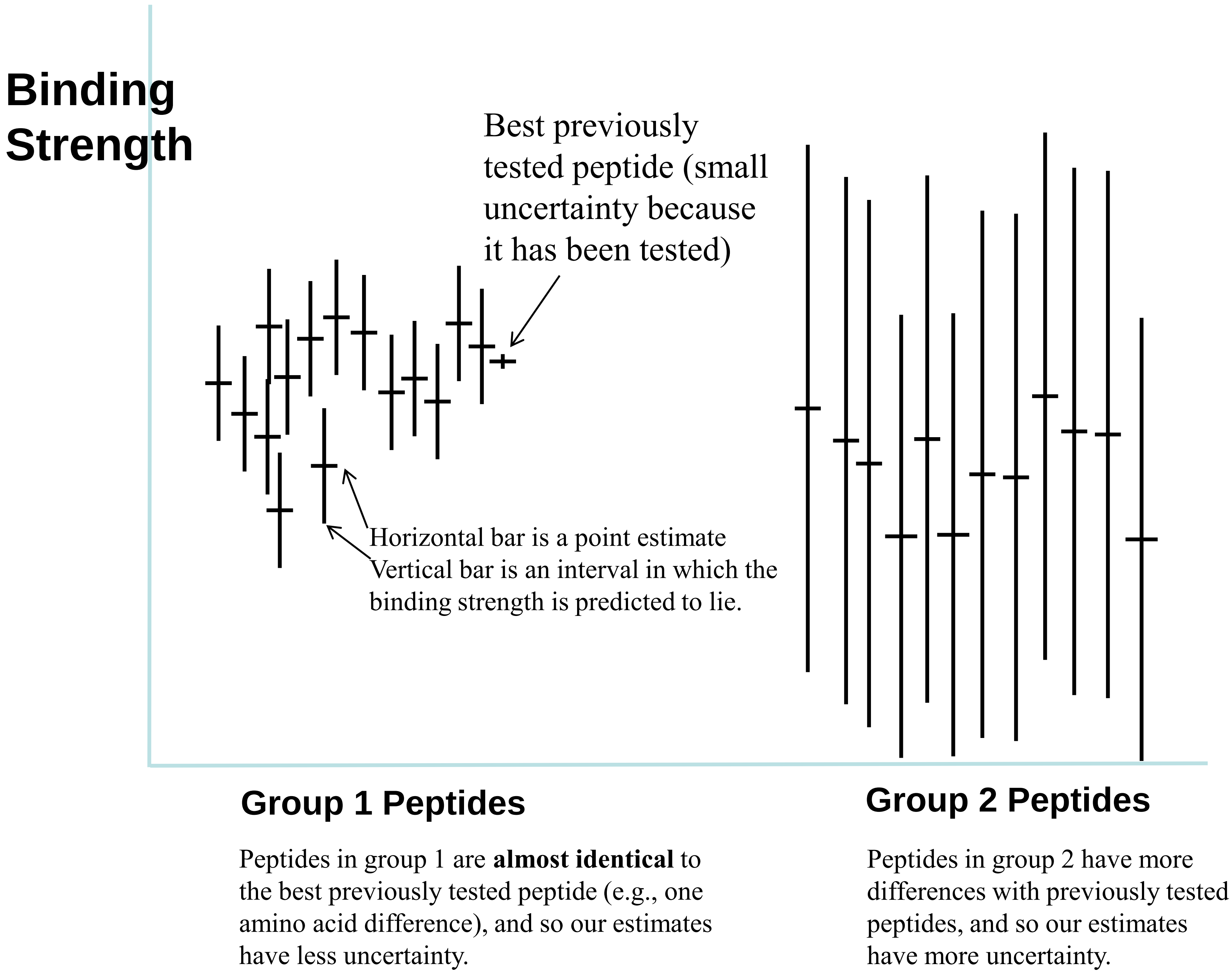


Collaboration with the Prasad Team

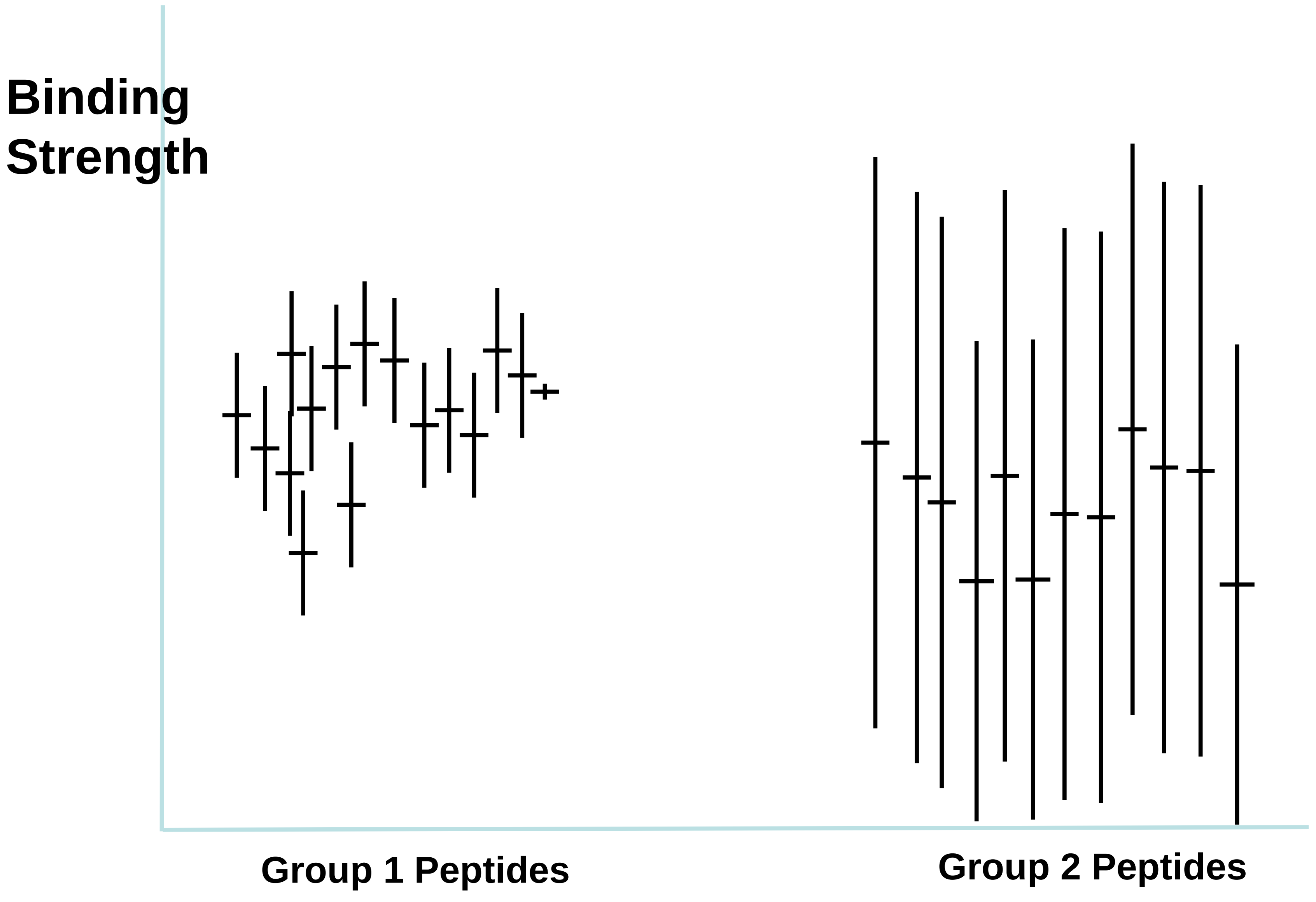
- Construction of the Bayesian prior:
  - To estimate binding against target materials, we are using a Bayesian linear model.
  - The set of features used include indicators of residues and motifs identified as important for binding against the target material via molecular dynamics simulations based on replica exchange.
  - We will fit prior distributions on the unknown parameters of the model from experimental data from other materials (quartz and carbon nano-materials).

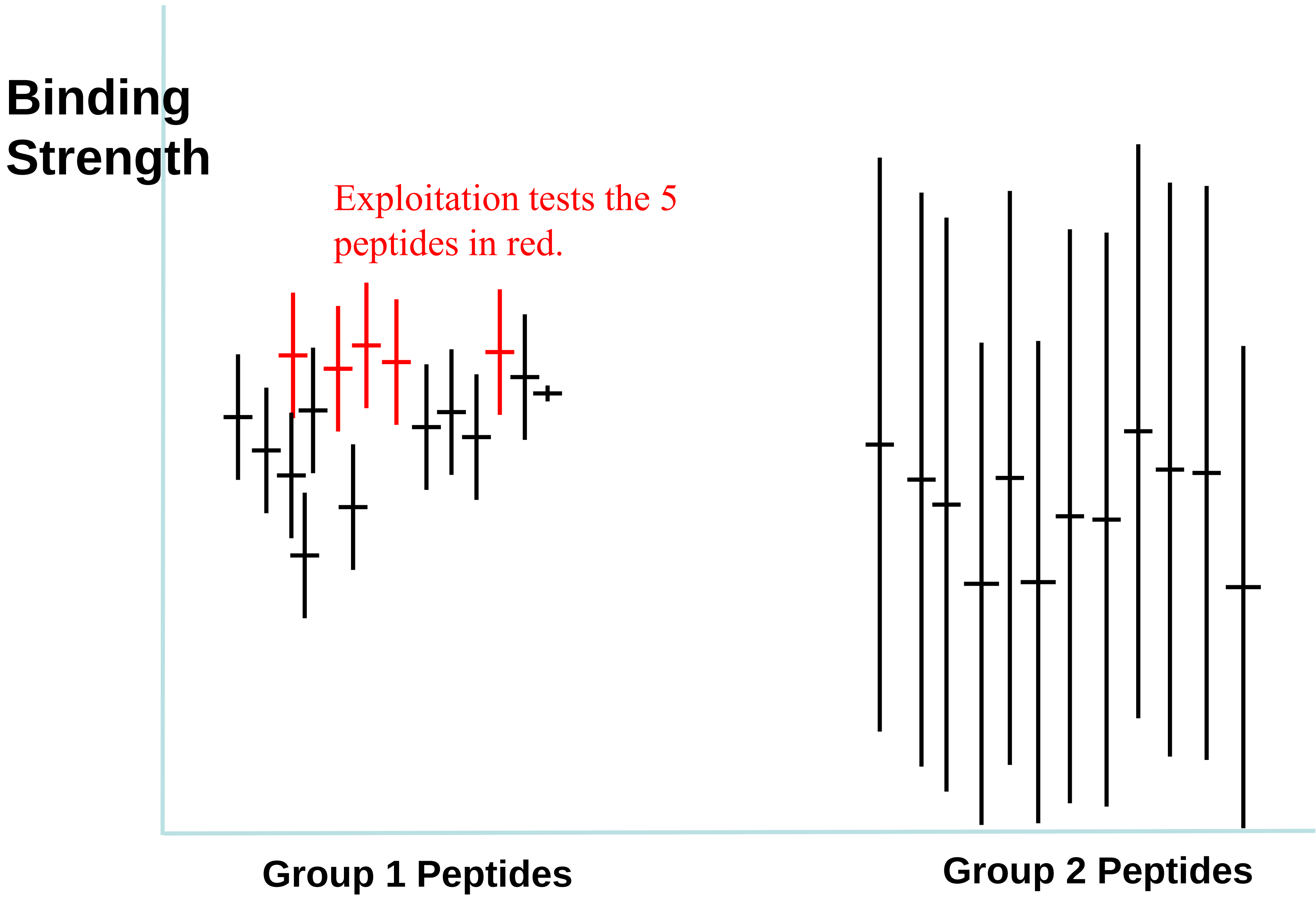
The value of optimal learning in peptide design

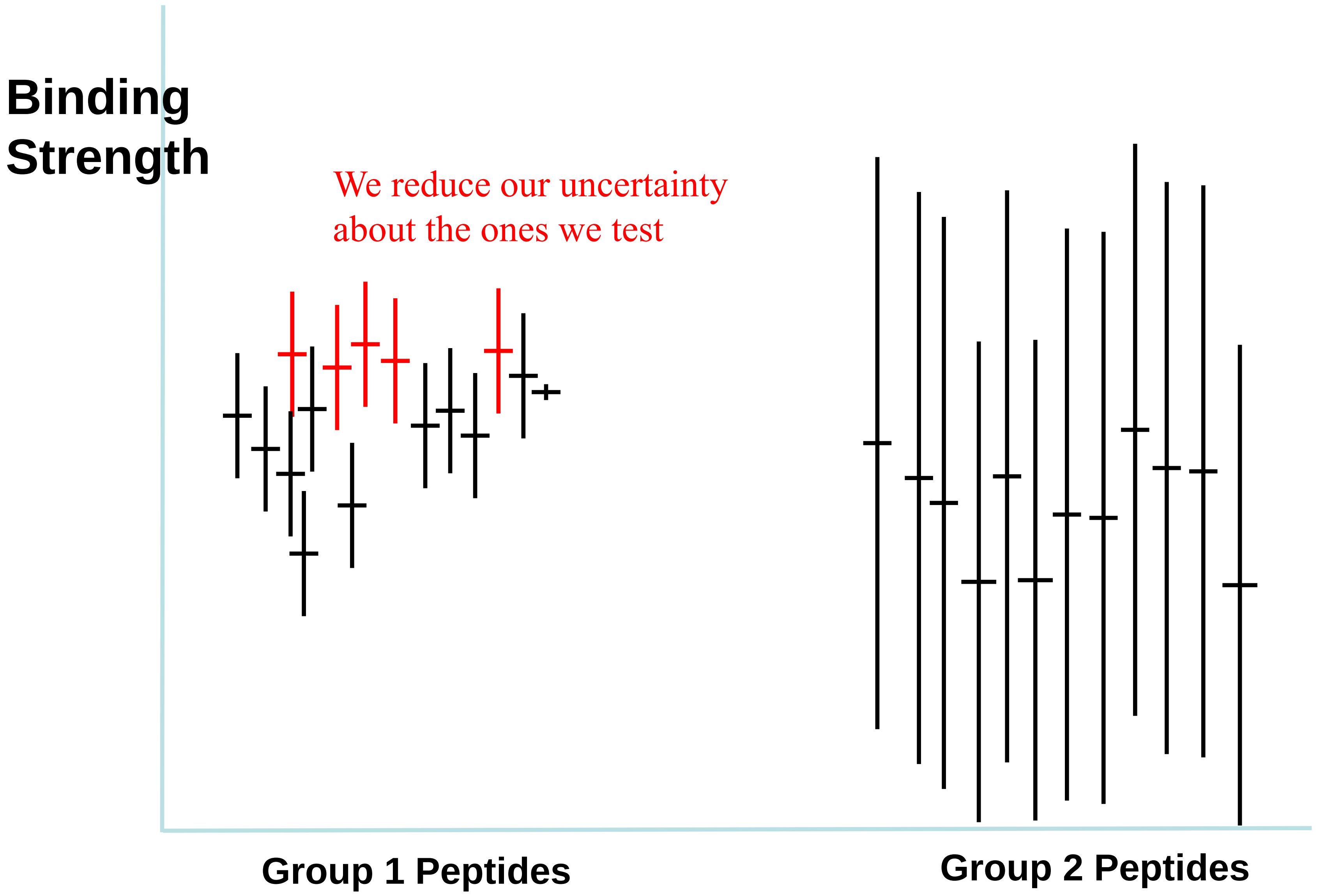
- Example showing why optimal learning is beneficial in peptide design.
  - Suppose we want to find a peptide with strong binding to a given target material.
  - We have identified a few peptides as binders through evolutionary search, and want to use this data to find ones that bind even better.
  - Let's compare two approaches:
    - 1. Use a statistical method to infer binding from the available data, select the top 10, and test these. (the "test the best" or "exploitation" strategy)
    - 2. Use optimal learning together with the same statistical method.

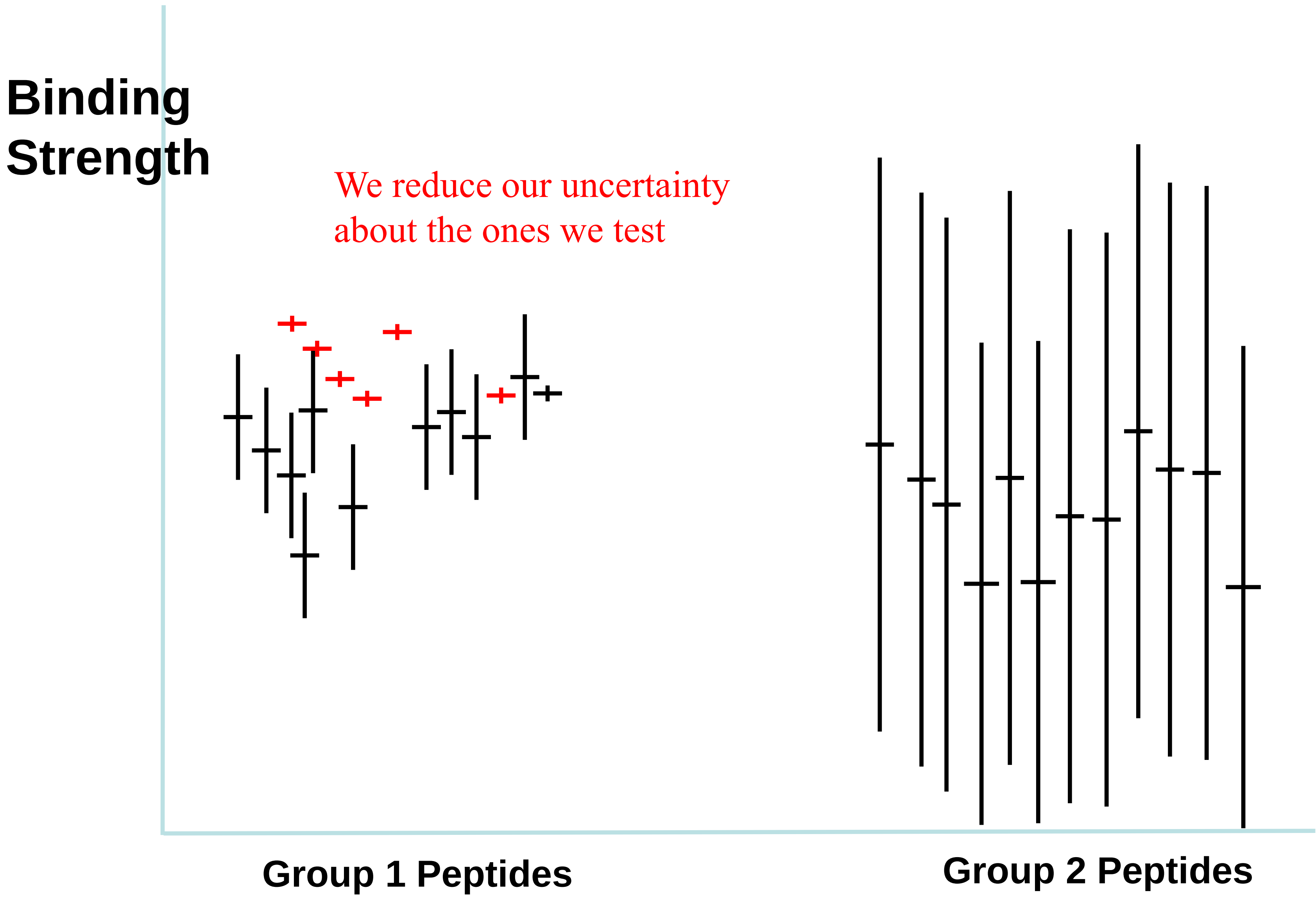


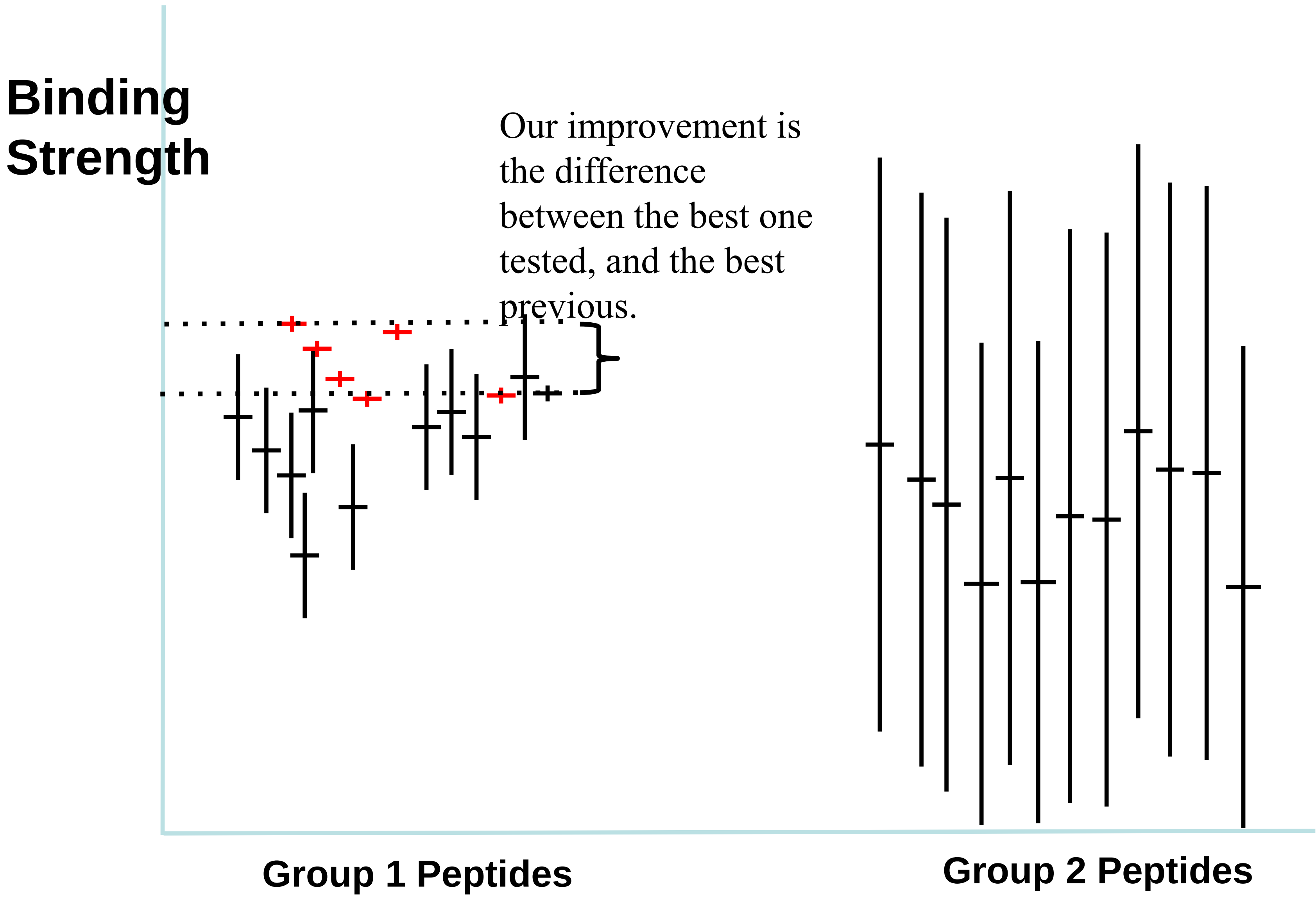
Binding
Strength
Best previously
tested peptide (small
uncertainty because
it has been tested)
Horizontal bar is a point estimate
Vertical bar is an interval in which the
binding strength is predicted to lie.
Group 1 Peptides
Peptides in group 1 are **almost identical** to
the best previously tested peptide (e.g., one
amino acid difference), and so our estimates
have less uncertainty.
Group 2 Peptides
Peptides in group 2 have more
differences with previously tested
peptides, and so our estimates
have more uncertainty.


Binding
Strength
Group 1 Peptides
Group 2 Peptides


Binding
Strength
Exploitation tests the 5
peptides in red.
Group 1 Peptides
Group 2 Peptides

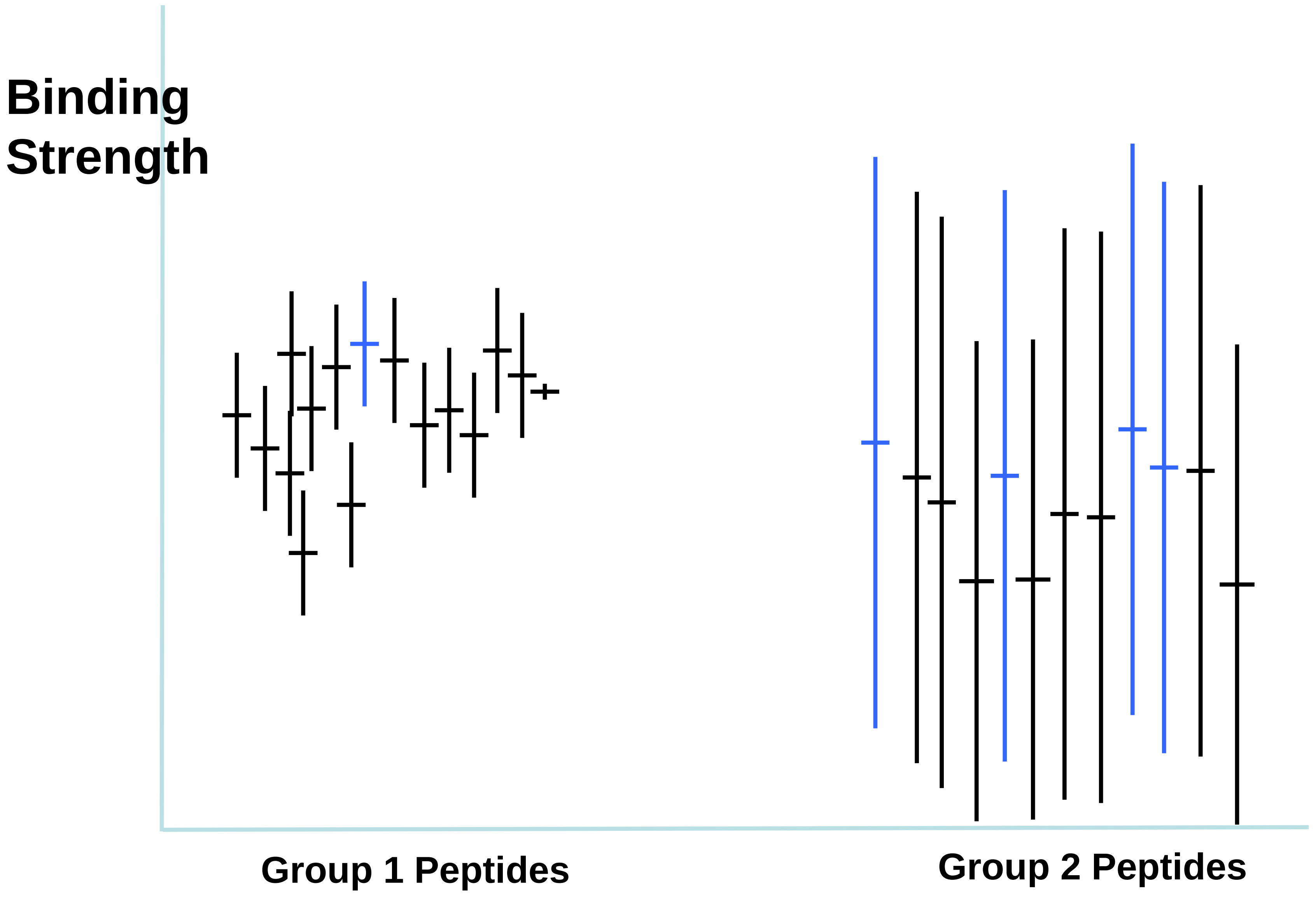

Binding Strength
We reduce our uncertainty about the ones we test
Group 1 Peptides
Group 2 Peptides

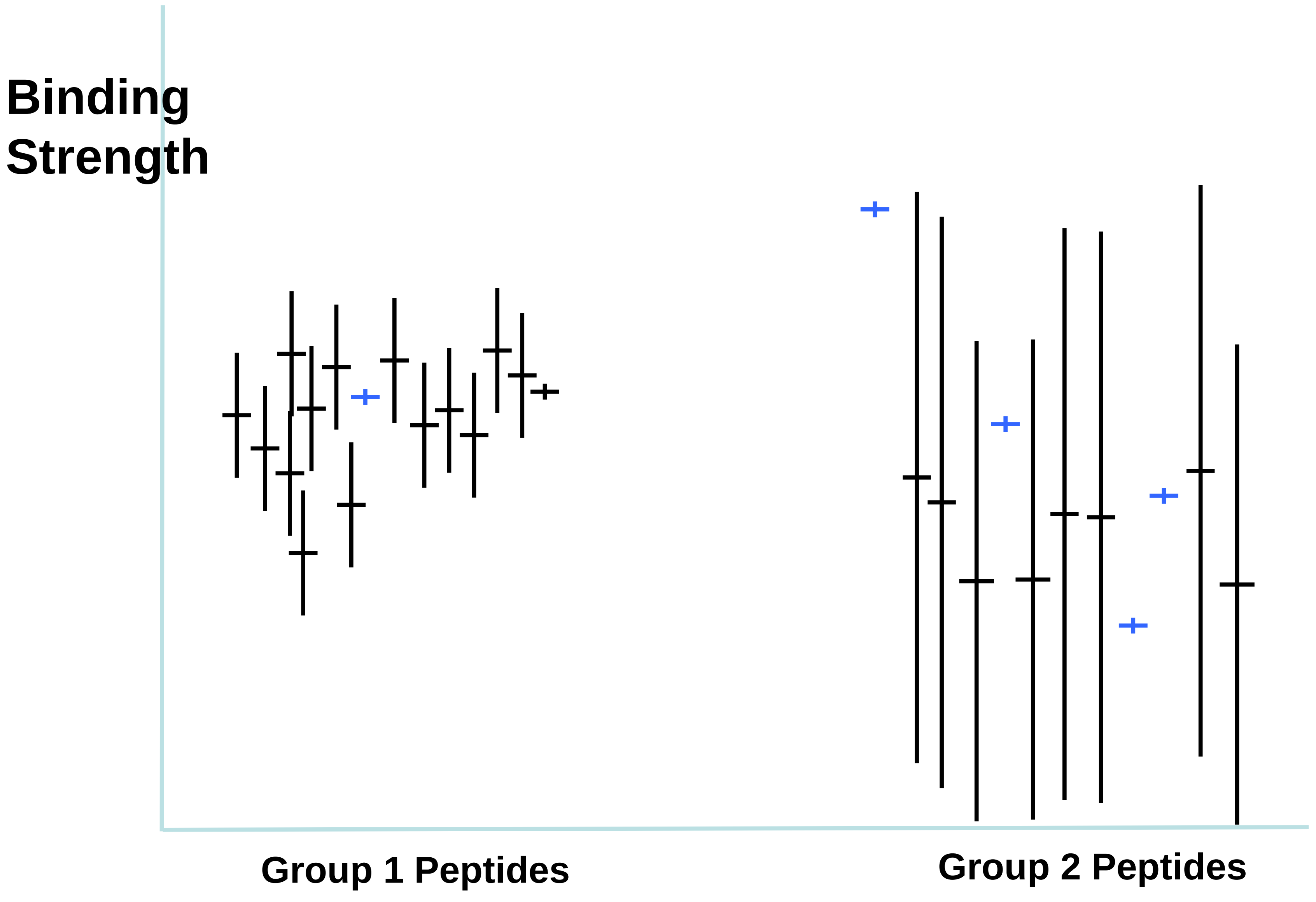

Binding
Strength
We reduce our uncertainty
about the ones we test
Group 1 Peptides
Group 2 Peptides

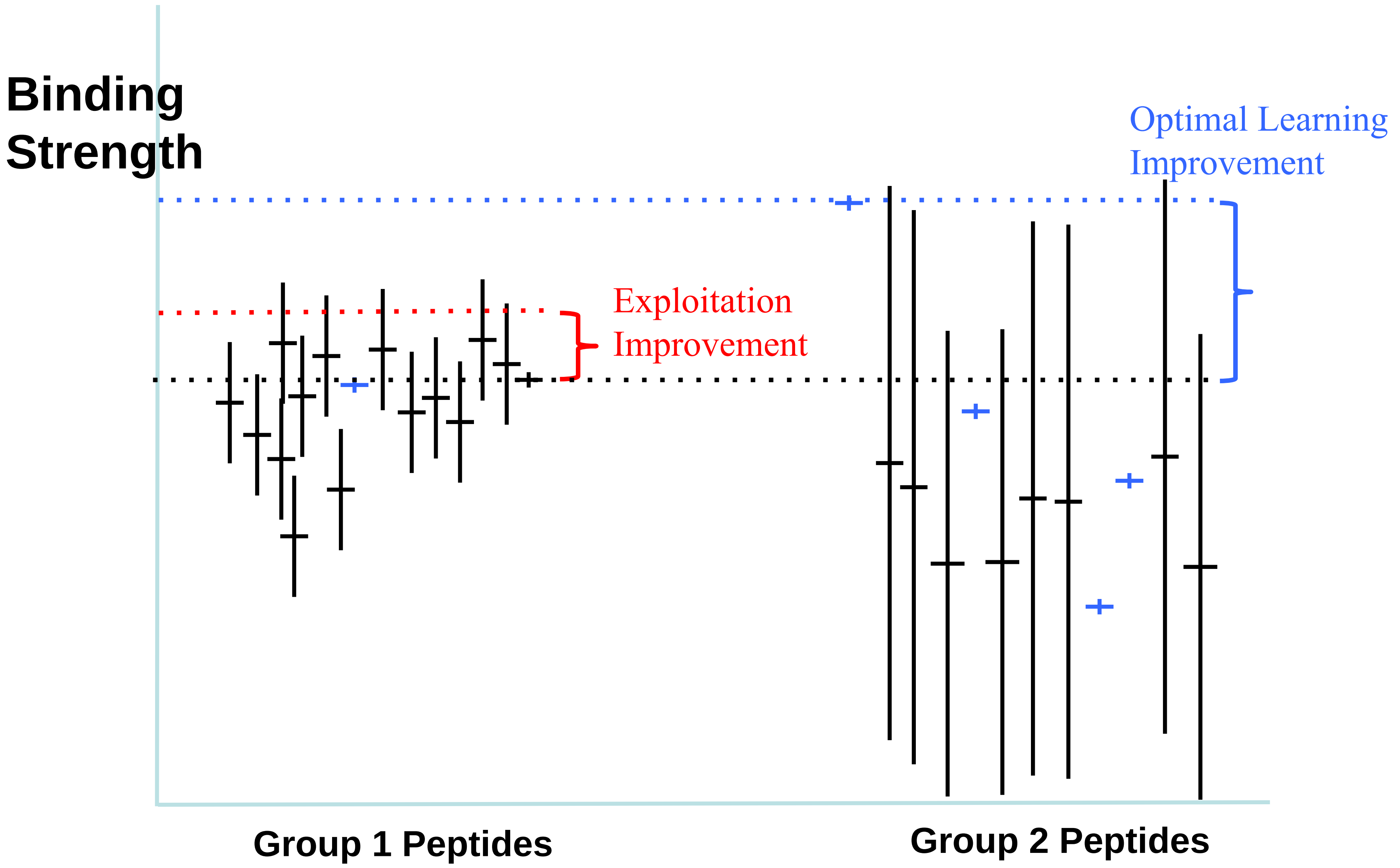

Binding
Strength
Our improvement is the difference between the best one tested, and the best previous.
Group 1 Peptides
Group 2 Peptides

The value of optimal learning in peptide design

- Instead suppose we use an optimal learning rule, which understands that we want to test peptides that have a good balance between:
  - Having large estimates,
  - Having high uncertainty, i.e., being different from what we’ve previously tested.
- This rule also understands correlations: closely related peptides have similar values, and so it is a good idea to spread measurements across different groups, in case one group is substantially worse than we thought.


Binding
Strength
Group 1 Peptides
Group 2 Peptides


Binding
Strength
Group 1 Peptides
Group 2 Peptides



Binding Strength
Optimal Learning Improvement
Exploitation Improvement
Group 1 Peptides
Group 2 Peptides

The value of optimal learning in peptide design

- The exploitation strategy worked poorly because we started in a situation where the peptides with best estimated binding energies also had low uncertainty.

- We do not always start in such a situation…
- …However, even if we don’t start there, if we use the exploitation strategy sequentially, we will soon reach this situation.

The value of optimal learning in peptide design

- By not explicitly pushing toward new areas of the search space, the exploitation strategy was only able to find an incremental improvement

- By investigating several high-risk high-reward options, the optimal learning strategy was able to obtain a bigger improvement.

Working with us

- We would like to work with you! To learn about your problem, we need to know:
  - What performance metric(s) are you trying to optimize?
    - Binding energy, lifetime, strength, melting temperature, …
  - What parameters affect these performance metric(s)?
    - Continuous parameters – Temperature, density, nano-particle size…
    - Discrete choices – Choice of amino acid, small molecule, substrate…
  - What experiments can you perform?
    - What data do they produce? How “expensive” are they to perform?
  - What do you already know about your process?
    - Your judgment and expertise.
    - Previously collected data, on this or similar problems.