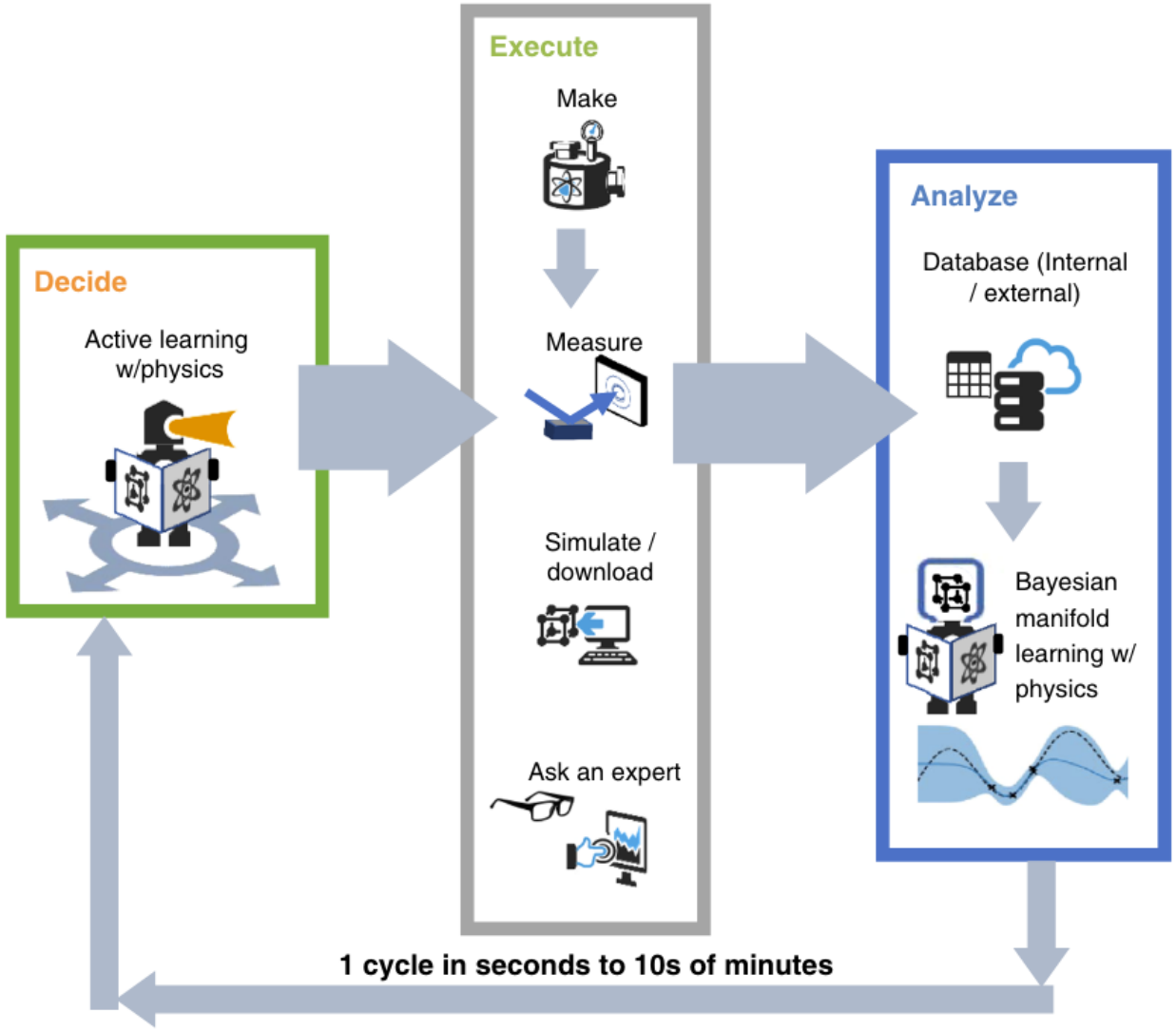
Real word examples of using Active Learning in Materials design and discovery

Lecture 19, Feb 22

On-the-fly closed-loop materials discovery via Bayesian active learning



- an autonomous materials discovery methodology for functional inorganic compounds which allows scientists to **fail smarter, learn faster**, and **spend fewer resources** in their studies

- CAMEO is implemented at the synchrotron beamline to accelerate the **interconnected tasks of phase mapping and property optimization**

10.1038/s41467-020-19597-w

The materials problem

- Explore the Ge–Sb–Te ternary system to identify an optimal phase-change memory (PCM) material for photonic switching devices
- have been used in DVD-RAM and nonvolatile phase-change random-access memory.
- find a compound with the highest optical contrast between amorphous and crystalline states in order to realize multi-level optical switching with a high signal-to-noise ratio.
- CAMEO is tasked to find the **composition** with **the largest difference** in the optical bandgap **ΔEg** and hence optical contrast between amorphous and crystalline states.

CAMEO Algorithm

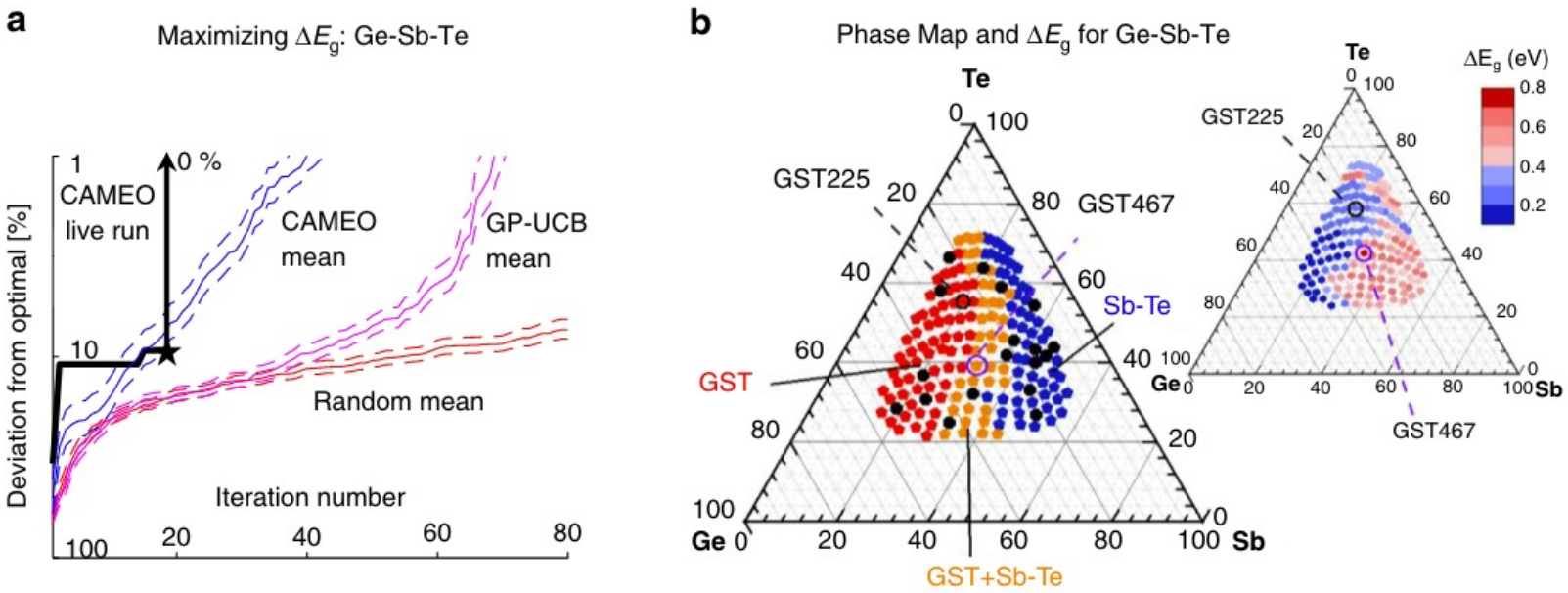
- First tries to figure out a phase map then switches to find optimal compositions maximizing the band gap

$$g(\mathbf{x}) = \begin{cases} P(\mathbf{x}), & c < 80\% \\ F(\mathbf{x}_r) = \mu(\mathbf{x}_r) + \beta\sigma(\mathbf{x}_r) + \gamma d(bfx_r), & else \end{cases}$$

- Optimization balances exploitation and exploration through the mean μ(xr) and weighted variance βσ(xr) much like the UCB algorithm

- The optimization acquisition function also allows the user to target points closer or further from phase boundaries via γd(xr), where d(xr)is the distance from point xr to the nearest phase boundary and γ is a user-defined parameter—negative (positive) to emphasize points near the edge (center) of the phase region.

CAMEO Algorithm



- A phase map is learned and fine-tuned using active learning

- Black star – iteration where a known optimal was found using the algorithm; rest are mean and std over 100 runs showing the CAMEO algorithm outperforms the optimization wrto UCB and random mean

Active Search

- we seek to sequentially inspect data to discover as many members of a desired class as possible with a limited budget
- The identities of the targets are unknown a priori but can be determined by querying an expensive oracle that can compute a label
- Given a budget T on the number of queries we can provide the oracle, we wish to design a policy that sequentially queries items to maximize the number of targets identified

https://www.youtube.com/watch?v=9y1HNY95LzY&ab_channel=ShaliJiang

A rough explanation of utility

- Given locations X and a label to denote whether something is a target in Y, we can define utility to be the number of targets found u(Y)
- When we maintain a probabilistic distribution for where the target locations can be found, we can “estimate” the expected utility

$$\mathbb{E}\left[u(\mathcal{D}_t \setminus \mathcal{D}_{t-1}) \mid X, \mathcal{D}_{t-1}\right] = \mathbb{E}_{Y|X,\mathcal{D}_{t-1}}\left[u(Y)\right] = \sum_{x \in X} \Pr(y = 1 \mid x, \mathcal{D}_{t-1}),$$

When only one iteration is left, it is best to choose a location with a high likelihood of being a target based on the posterior

$$\mathbb{E}\left[u(\mathcal{D}_t \setminus \mathcal{D}_i) \mid X, \mathcal{D}_i\right] = \sum_{x \in X} \Pr(y = 1 \mid x, \mathcal{D}_i) + \mathbb{E}_{Y|X,\mathcal{D}_i}\left[\max_{X'} \mathbb{E}\left[u(\mathcal{D}_t \setminus \mathcal{D}_{i+1}) \mid X', \mathcal{D}_{i+1}\right]\right],$$

The above thinking can be extended using what is called a Bellman’s equation

Application to finding bulk metallic glasses

- The goal here is to find novel alloys capable of forming bulk metallic glasses (BMGs).
- Compared to crystalline alloys, BMGs have many desirable properties, including high toughness and good wear resistance.
- This dataset consists of 118 678 known alloys from the materials literature among which 4 746 (about 4%) are known to exhibit glass-forming ability, which we define as positive/targets.

- Or in **virtual screening** for drug discovery -- of a large database of compounds searching for those that show binding activity against some biological target.

T-test based evaluation of the proposed method

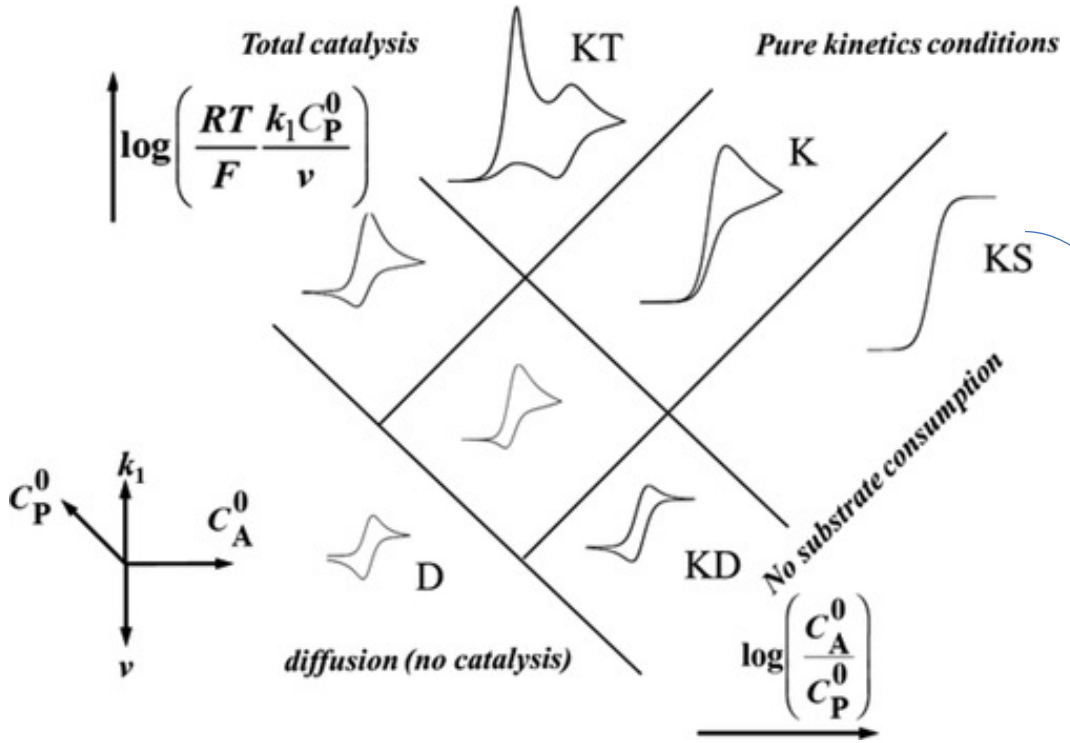
Table 3: Results for 10 drug discovery datasets in batch setting: Average number of positive compounds found by the baseline *uncertain-greedy* batch, greedy-batch, sequential simulation and batch-ENS policies. Each column corresponds to a batch size, and each row a policy. Each entry is an average over 200 experiments (10 datasets by 20 experiments). The budget T is 500. Highlighted are the best (bold) for each batch size and those that are not significantly worse (blue italic) than the best under one-sided paired t-tests with significance level $\alpha = 0.05$.

Different decision policies

	1	5	10	15	20	25	50	75	100
UGB	-	257.6	257.9	258.3	250.1	246.0	218.8	206.2	172.1
greedy	269.8	268.1	264.1	261.6	258.2	257.0	240.1	227.2	208.2
ss-one-1	269.8	260.7	254.6	245.2	233.6	223.4	200.8	182.9	178.9
ss-one-m	269.8	264.5	257.7	250.0	244.4	236.5	211.7	195.4	179.4
ss-one-s	269.8	266.8	261.3	256.7	248.7	244.1	214.9	202.4	181.3
ss-one-0	269.8	268.1	264.1	261.6	258.2	257.0	240.1	227.2	208.2
ss-two-1	281.1	237.1	219.8	210.8	212.1	196.2	172.1	158.8	152.9
ss-two-m	281.1	252.6	246.4	237.2	232.9	225.1	200.2	181.6	167.2
ss-two-s	281.1	248.9	242.5	235.3	226.6	219.2	196.7	175.3	158.3
ss-two-0	281.1	252.5	247.6	247.9	244.4	240.4	225.6	213.8	199.1
ss-ENS-1	**295.1**	269.4	247.9	227.2	223.1	210.3	185.3	152.6	148.7
ss-ENS-m	*295.1*	293.8	290.2	285.3	281.6	274.4	249.4	217.2	203.1
ss-ENS-s	*295.1*	289.9	278.3	269.8	262.6	255.0	220.8	185.5	161.2
ss-ENS-0	*295.1*	293.6	289.1	288.1	*287.5*	280.7	269.2	257.2	241.0
batch-ENS-16	*295.1*	**300.8**	**296.2**	293.9	**292.1**	*288.0*	275.8	*272.3*	252.9
batch-ENS-32	*295.1*	*300.8*	*295.5*	**297.9**	*290.6*	**288.8**	**281.4**	**275.5**	**263.5**

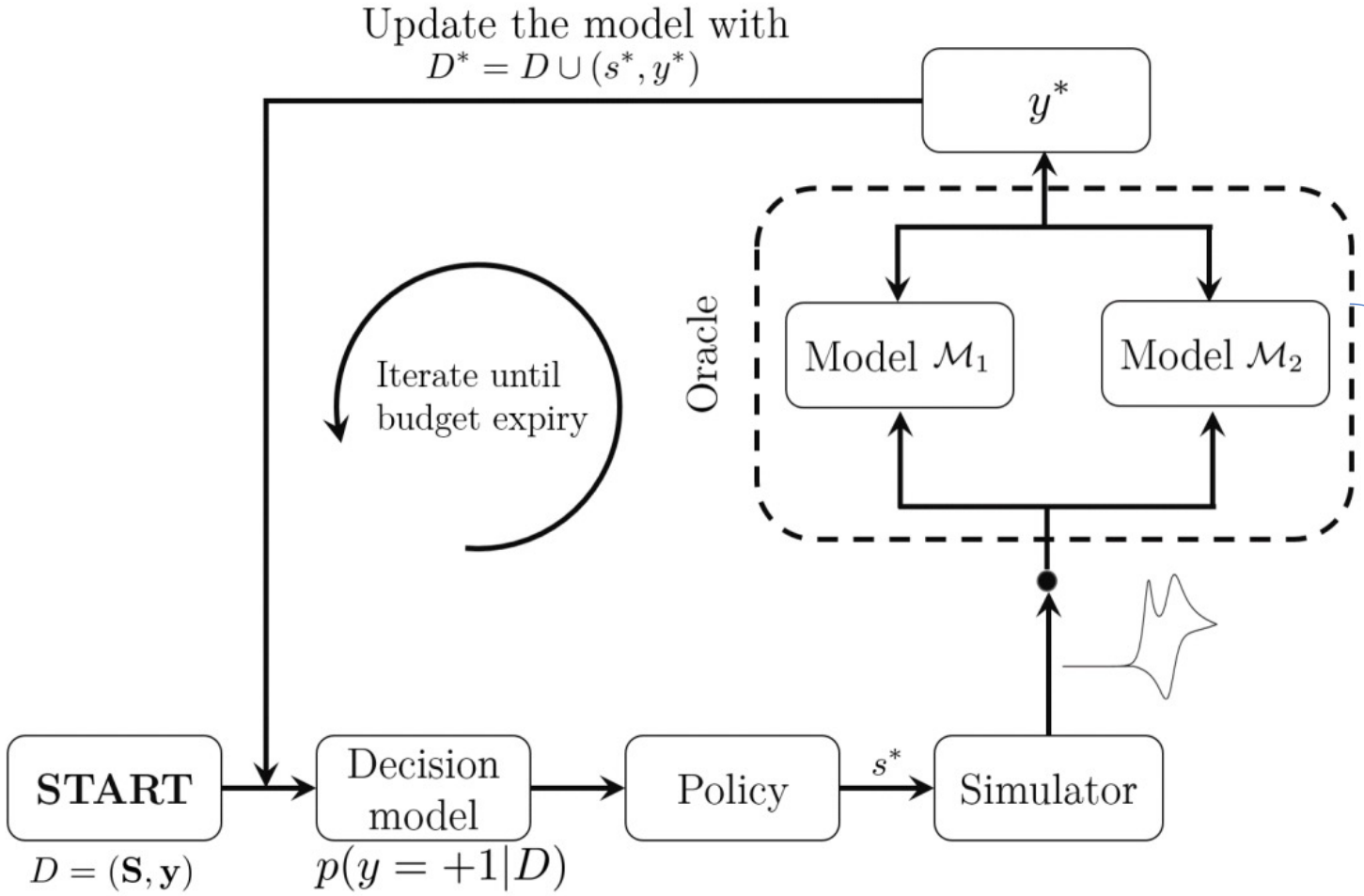
Application to data-driven discovery of bifunctional catalysts

The goal is to find catalyst(s) that can work the best in both Oxygen evolution and reduction reaction for Hydrogen based or Metal-air batteries



Roughly, any catalysts that result in this shape of a CV curve are desired as it is a signature of a high catalytic activity

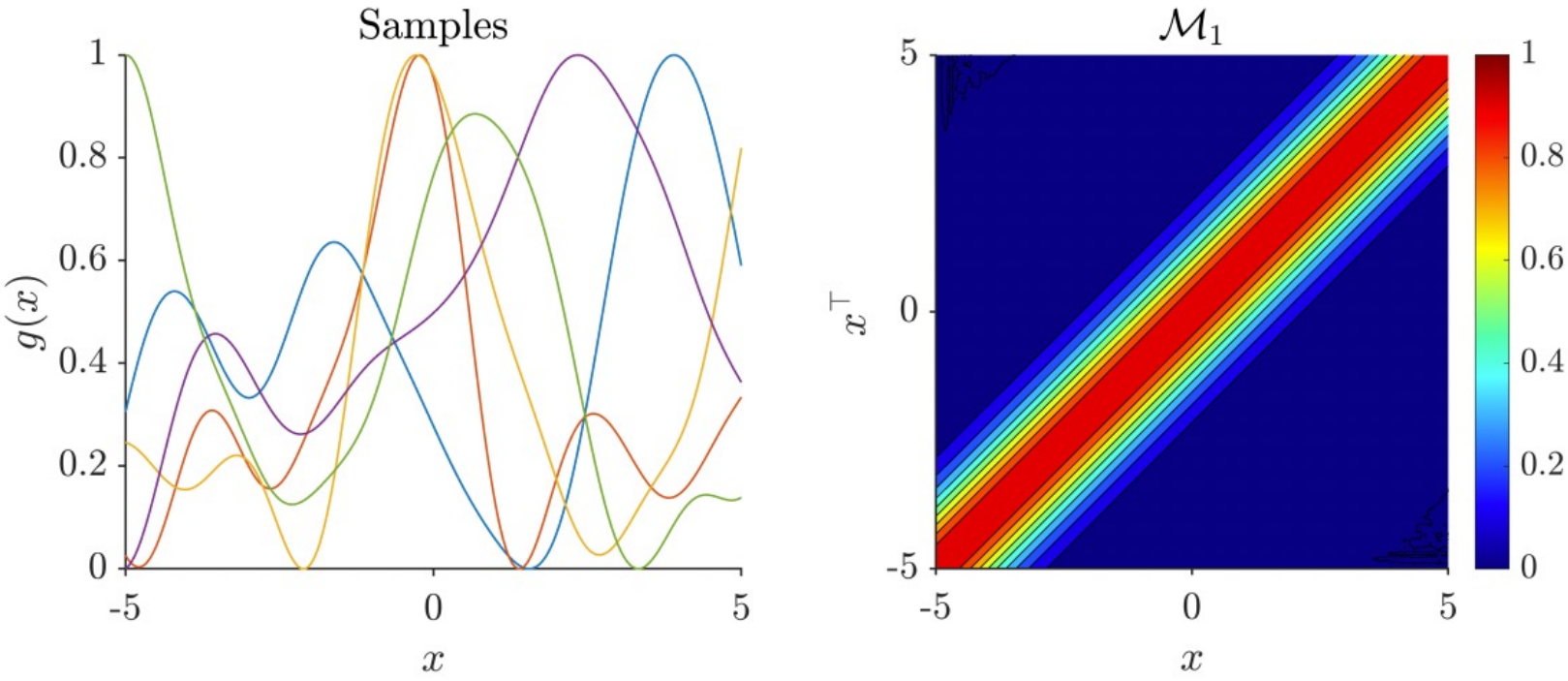
Application to data-driven discovery of bifunctional catalysts



Fortunately, it is easier to differentiate the shapes based on specific covariances in a GP using a Bayesian Model Selection approach

https://doi.org/10.26434/chemrxiv.14569035.v1

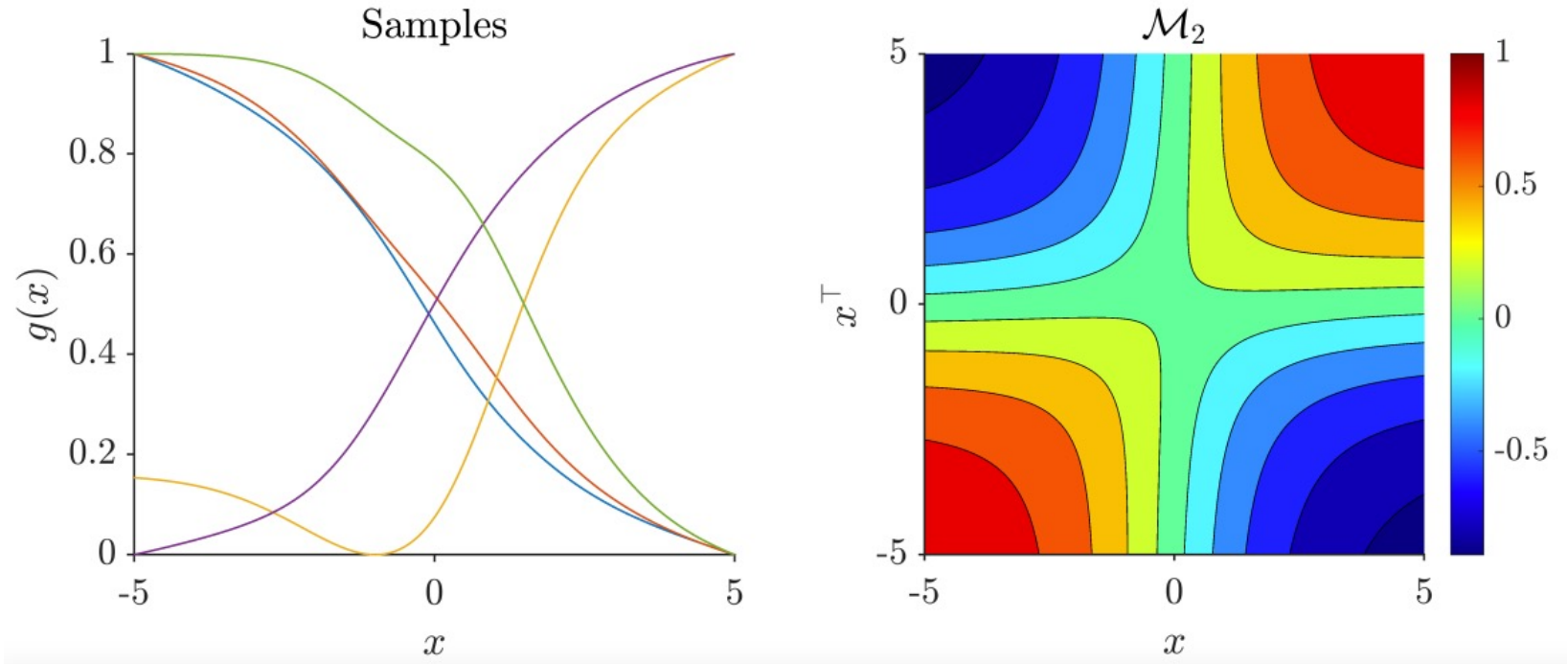
Null covariance for all CV's



$$k(x, x') = \sigma_f^2 \exp\left((x - x')^\top \Lambda^{-1} (x - x')\right)$$

We represent each CV curve as a function of time and voltage thus x has two dimensions.

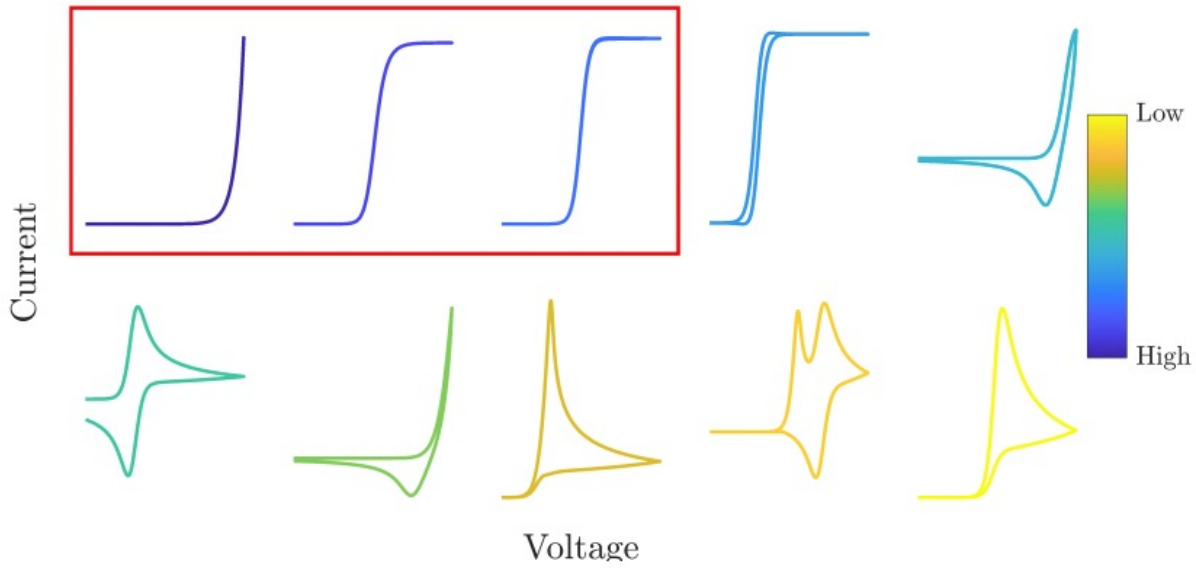
covariance for S-shaped CV's

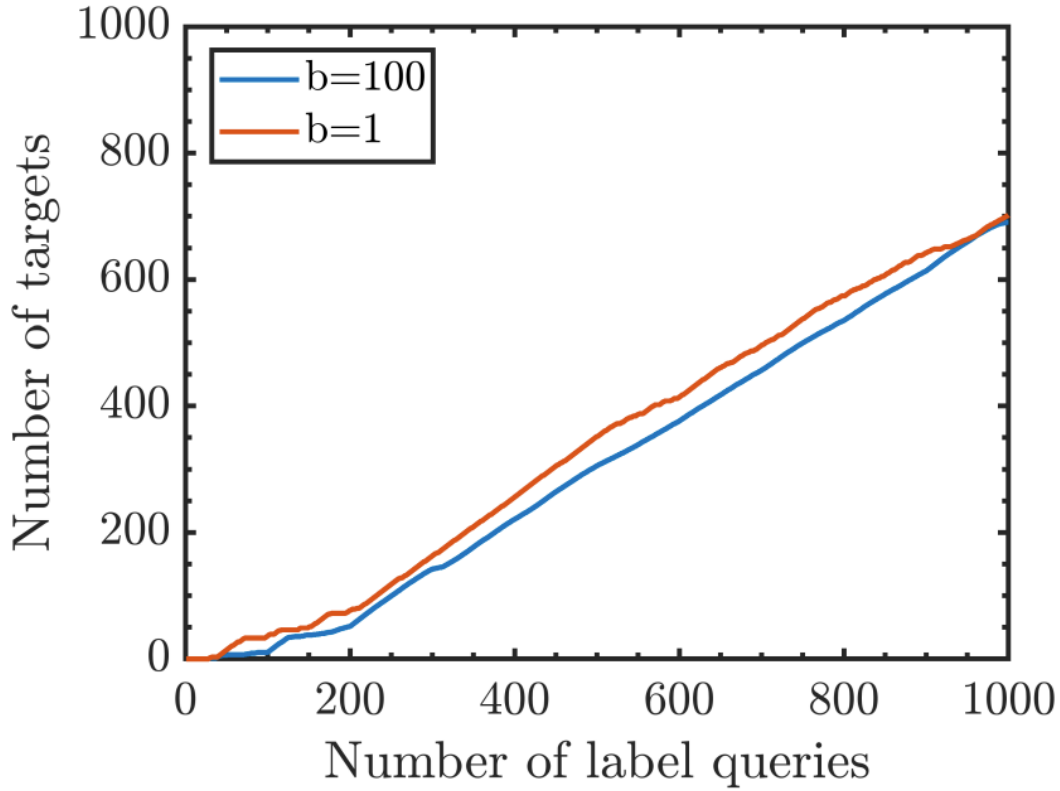


$$k(x, x') = \sigma_f^2 \sin^{-1}\left(\frac{x^\top \Lambda^{-2} x'}{\sqrt{h(x)h(x')}}\right)$$

$$h(x) = 1 + x^\top \Lambda^{-2} x$$

covariance for S-shaped CV's





1000 labels roughly correspond to 6% of the total possible query locations where the targets are less than 1%