Active Knowledge Extraction from Cyclic Voltammetry

Kiran Vaddi* and Olga Wodo*

Department of Materials Design and Innovation, University at Buffalo, Buffalo, NY, USA

E-mail: kiranvad@buffalo.edu; olgawodo@buffalo.edu

Phone: +1 (716)645 1377

Abstract

Cyclic Voltammetry (CV) is an electro-chemical characterization technique used in an initial material screening for desired properties and to extract information about electro-chemical reactions. In some applications, to extract kinetic information of the associated reactions (e.g., rate constants and turn over frequencies), CV curve should have a specific shape (for example an S-shape). However, often the settings to obtain such curve are not known *a priori*. In this paper, an active search framework is defined to accelerate identification of settings that enable knowledge extraction from CV experiments. Towards this goal, a function space representation of CV responses is used in combination with Bayesian Model Selection (BMS) method to efficiently label the response to be either *S-shape* or not *S-shape*. Using an active search with BMS oracle, we report a linear target identification in a 6-dimensional design space (comprising of thermodynamic, mass transfer and solution variables as dimensions). Our framework has the potential to be a powerful virtual screening technique for molecular catalysts, bi-functional fuel cell catalysts etc.

Keywords

Accelerated Catalyst Discovery, Gaussian Processes, Bayesian Model Selection, Active Learning, Cyclic Voltammetry.

Introduction

Cyclic Voltammetry (CV) is an electro-chemical characterization technique that measures current generated under a cyclic voltage load between a initial and final voltage varied at a given rate. The measured current is a highly non-linear response from various physical phenomenon such as mass transport, kinetics, adsorption etc. In principle, it is possible to determine the properties associated with the underlying physical phenomenon. However, the property extraction is a non-trivial task. In a CV experiment, a steady state current is obtained when all reactions in the mechanism have the same apparent rate constants [1]. This is because the facile reactions in the sequence are held back from their maximum rates by the sluggish reactions called a *rate determining step* that also determines the magnitude of steady state current. Extracting rate constants of the rate determining step thus requires the CV curve to be in a S-shape [3, 21] with a clear steady state current region resolved during measurement. Towards this goal, obtaining a S-shaped CV curve requires the experiment to be run with a set of conditions (e.g, temperature, substrate concentration, scan rate), amenable for S-shape CV curves which are unknown *a priori*. Moreover, choosing conditions where a given electrochemical system exhibits a S-shaped CV curve is dependent on underlying system of electrochemical reaction(s) which is(are) also unknown for novel materials. In the absence of a known mechanism, an exhaustive search over all the possible tunable parameters is performed [16] to narrow down the region of interest. Such exhaustive strategy comes at a price of very high computational cost especially in a high-dimensional search space of multiple complex reaction mechanisms.

As an alternative approach, experts define a figure of merit (FOM) [21](a performance measure) as a proxy signature of a physical phenomenon of interest. FOM extracted from CV can be also used in material discovery using data-driven methods. For example, in [23, 15, 9, 24] different types of FOM have been used for catalyst discovery using data-driven methods. With FOM defined, the goal is to find a material that produces a response with a FOM that is better than that of known materials. For instance, in case of a high-throughput exploration for a new catalyst, the overpotential is a common FOM [18] (or performance measure) used in the combinatorial searches [24]. The over-potential can be thought of as the voltage (beyond the thermodynamic requirement) required to produce a (pre-defined) target current. This FOM has clear utility to screen for well performing materials, but misses on the main advantage of CV - that is the capability to extract the kinetic information (such as rate constants [25], turn-over frequencies [4, 17]).

Given the time and financial constraints, we propose to accelerate the process of extracting kinetic information from CV curves using the active learning technique [10, 19, 6]. Rather than relying on the selection of figure of merit, we build function space representations of our target (S-shaped) and non-target (everything else) CV responses and use Bayesian Model Selection (BMS) for automatic classification. We encode prior knowledge of target and non-target CV responses using the basis functions of a function space representation using Gaussian processes ($\mathcal{GP}$). $\mathcal{GP}$ have been previously used to infer the kinetic parameters [8, 20] of a CV response by using a maximum likelihood estimate and $\mathcal{GP}$ regression. In another work [15], a Bayesian approach is used to search for an approximate rate constant when the reaction mechanism is known. In this work, however, we use $\mathcal{GP}$ as a data representation model to distinguish S-shaped CV curves from other types of continuous CV curves. Once a S-shaped CV curve is collected, the foot-of-the-wave analysis (FOWA) [25] can be used to extract the rate constant of a rate determining step. When combined together with FOWA, the proposed approach can be a robust technique that does not require any knowledge of the actual reaction mechanism.

In this work, we focus on S-shaped CV responses due to their utility for: a) extracting kinetic information [3, 21]–using the foot of the wave analysis [25] that can only be applied to a S-shaped CV curve. b) screening for bi-functional catalysts – materials that produce CV curves similar to S-shape in two different voltage sweep ranges [2, 11]. While the two applications are different, they can be approached under a common framework of active search in a combinatorial space, where we are interested in finding S-shaped CV curves within the combinatorial space.

The rest of the paper is organized as follows: (i) First we introduce Bayesian active learning framework with a general probabilistic model. We establish a connection between collected data at observed locations with the oracle used to classify and update the decision model used for active learning. (ii) We then introduce the Bayesian Model Selection (BMS) procedure that computes a classification preference for targets and non-targets based on collected data and set of parametric models. (iii) We introduce a $\mathcal{GP}$ model that builds a function space representation for collected data to use as a parametric model in BMS. (iv) We apply our methodology on a search space of a simple EC mechanism and demonstrate the application of the BMS oracle to classify CV responses in order of its S-shape. (v) Finally, we use the BMS oracle in active search to address the challenges in knowledge extraction, virtual screening of materials for electrochemical applications using cyclic voltammetry.

Methods

Our goal is to identify the measurement settings from which one can extract kinetic information captured in a CV response. Towards this goal, we seek to identify measurement conditions for which an S-shape CV curve is collected and registered as such by our oracle. We use an active learning technique summarized in Figure 1 to accelerate the search for measurement settings within fixed computational budget. Our active learning approach involves iterative collection of data points from a search space $\mathcal{S}$. The process starts with a small set of observed data $D = (\mathbf{S}, \mathbf{Y})$ where $\mathbf{S} \in \mathcal{S}$ are the observed locations and $\mathbf{Y} \in \{-1, 1\}$ are corresponding labels. In each iteration, the algorithm collects data and incrementally updates the decision model $p(y = +1|D)$ it aims to learn with y representing a label. A user-defined selector (or policy) identifies a (or a batch of) candidate location(s) in the search space for observing the responses. The policy typically maximizes a utility function given the decision model. For example, given D, we can define a policy using a utility function that simply counts number of targets in the dataset $u(\mathbf{S}) = \sum\limits_{s_i \in \mathbf{S}, y_i \in \mathbf{Y}} [y_i = +1]$. A policy can be defined to potentially select more targets to be added to the data pool D using:

$$s^* = \underset{s}{\operatorname{argmax}}\, \mathbb{E}\left[u(\mathcal{S} \setminus \mathbf{S}|D)\right] \tag{1}$$

Where $\mathbb{E}[.]$ represents expectation. Given a location $s^* \in \mathcal{S}$, the corresponding experiment is performed and a response is collected. In this work, we collect CV response curve from a CV curve simulator and the response is then passed to an oracle. Oracle labels the response to be either a target or non-target (for example in Figure 1, we show a non-target like CV shape which will be assigned $y^* = -1$ as a label). The next step is to augment D using the data collected in the current iteration i.e. $D^* = D \cup (s^*, y^*)$. The decision model is then updated with D^*. This process is repeated until computational budget–defined in terms of total number of label queries or equivalently number of simulations–is exhausted. As an oracle we use Bayesian Model Selection (BMS) that operates on two models $\mathcal{M}_1, \mathcal{M}_2$ referred to as null model (representing a typical CV curve) and target model (representing an S-shaped CV curve), respectively. Moreover, we use a variation of active learning called active search [7] which maximizes the number of targets found in contrast to traditional active learning where the selector is defined with a goal to closely approximate $p(y = +1|D)$.

Bayesian Model Selection

The key component of our active learning framework is the oracle. We use BMS as a tool to identify a preferred model from a family of parametric probability distributions, each of

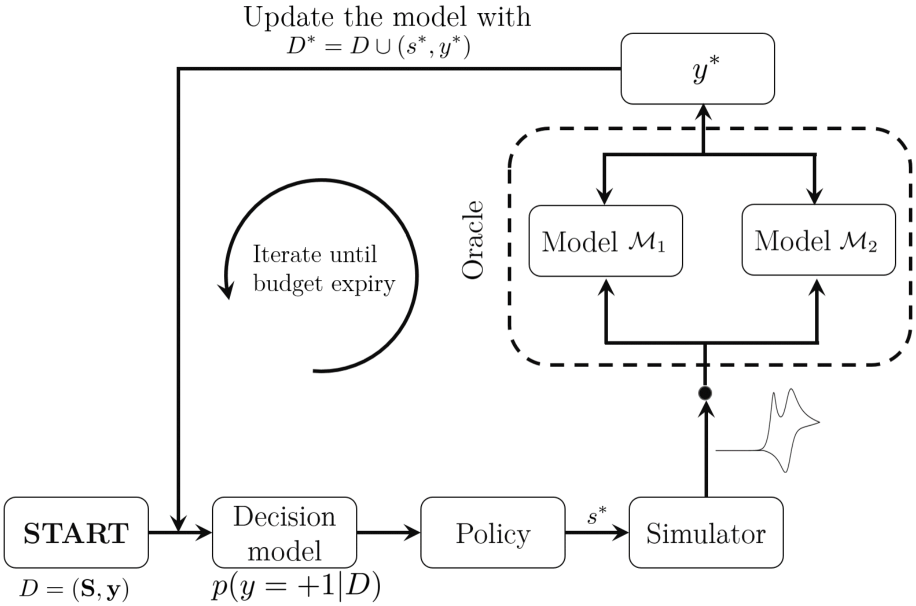

Figure 1: Active learning framework as a flowchart. Active learning iterations start with a few labelled data points in the search space $\mathcal{S}$. We stop collecting data when we have selected a pre-defined number (called budget) of locations for updating our decision (or belief) model.

which can explain the observed data with differing degrees of fidelity. Using a supervised learning procedure that compares an input $\mathbf{X}$ and output $\mathbf{y}$, we compute a model posterior using Bayes rule to select the model that best explains the observed data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$.

Here, given the observed data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, we compute probability that the data is sampled from any given model encoding our prior information. Computed posterior probabilities will be used as a score to differentiate whether the collected response (for example, a CV response at any input location in the search space of materials) is a target (with higher probability for the corresponding target model) or not. In this work, we use both BMS and active learning in a related but different context. BMS is used with observed data encoding a single CV curve while active learning is used in the search space with their corresponding binary labels (i.e. a target or not) as observed data. Moreover, BMS is used as an oracle for the active learning task with models $\mathcal{M}_j$ as $\mathcal{GP}$.

For each model $\mathcal{M}$ with a parameter index θ– a concatenated vector of hyper-parameters–

we first compute model evidence $p(\mathbf{y}|\mathbf{X}, \mathcal{M})$ on the observed data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$:

$$p(\mathbf{y}|\mathbf{X}, \mathcal{M}) = \int p(\mathbf{y}|\mathbf{X}, \theta, \mathcal{M}) p(\theta|\mathcal{M}) \, \mathrm{d}\theta \tag{2}$$

where $p(\mathbf{y}|\mathbf{X}, \theta, \mathcal{M})$ is probability of obtaining outputs $\mathbf{y}$ given input data $\mathbf{X}$ and a model $\mathcal{M}$. $p(\theta|\mathcal{M})$ represents distribution of parameter θ for any given model $\mathcal{M}$.

To understand which model to prefer from a finite set of models $\{\mathcal{M}_i\}_{i=1}^{n}$, we apply the Bayes rule to compute the posterior probability of each model $\mathcal{M}_j (j \in \{1, 2, ..n\})$ given data $\mathcal{D}$ using the posterior of Equation (2):

$$p(\mathcal{M}_j|\mathcal{D}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathcal{M}_j) p(\mathcal{M})}{p(\mathbf{y}, \mathbf{X})} \tag{3}$$

where $p(\mathcal{M})$ represents a prior over the finite set of models that is typically taken to be uniform i.e. no prior preference to any single model. One common approach is to use logarithm of the probability which can be interpreted as the information content of a probability model given data. Taking the logarithm of Equation (3), we get the following:

$$\log p(\mathcal{M}_j|\mathcal{D}) = -\log \left[1 + \sum_{i \neq j}^{n} \frac{p(\mathbf{y}|\mathbf{X}, \mathcal{M}_i)}{p(\mathbf{y}|\mathbf{X}, \mathcal{M}_j)} \right] \tag{4}$$

$\mathcal{GP}$ Models for Catalytic Responses

A $\mathcal{GP}$ is a distribution over smooth latent functions $g : \mathcal{X} \to \mathbb{R}$. Assuming the observation model $p(y|g)$ is known, the standard approach is to use non-parametric Bayesian approach by placing a $\mathcal{GP}$ distribution over g, i.e. $p(g) = \mathcal{GP}\big(\mu(x), k(x, x')\big)$. Here $\mu(x) : \mathcal{X} \to \mathbb{R}$ is a mean function and $k(x, x') : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a covariance function. A function-space viewpoint provides an intuitive explanation of $\mathcal{GP}$ as vector space of functions in a chosen (potentially non-linear) feature space with $\phi(x)$ as a basis. In the function space representation, the observation model plays the role of weights W with function g represented using $g(x) = \phi(x)^\top W$. It can be shown that $\phi(x)$ can be implicitly defined using the covariance function

$k(x, x')$ between pair of inputs $x, x' \in \mathcal{X}$ and a mean function $\mu(x)$ of W [1]. The mean function μ encodes an average behavior of the function g. The covariance function $k(x, x')$ encodes the correlations between outputs $g(x), g(x')$ for any given pair of input points (x, x'). In this work, we denote the concatenated vector of the parameters in $\mu(x)$ and $k(x, x')$ as θ. Once we select a $\mathcal{GP}$ encoding our prior beliefs, we use Bayes rule to update our posterior $p(g|\mathcal{D})$ conditioned on observed data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ where $\mathbf{y}$ are the discrete evaluations of function g at inputs $\mathbf{X}$. For more information on $\mathcal{GP}$, readers are referred to [26].

We represent a typical response from a cyclic voltammetry experiment as a function $I(t, v)$ with I being the current response collected at a time t for a time dependent applied voltage $v = V(t)$. The voltage load $V(t)$ is typically chosen to be linear and the voltammetry is often referred as direct current voltammetry [15]. Classification of a CV into an S-shape (or not S-shape) can be looked at as determining a model evidence of a function defined by CV curve $(v, t) \mapsto I$ under a $\mathcal{GP}$ function space with observed data $\mathcal{D}$ given by the discrete CV curve $\mathbf{X} = I, \mathbf{y} = (v, t)$. The covariance of a CV curve gives rise to the basis functions in the $\mathcal{GP}$ space and the time-voltage grid becomes the input space where the function is evaluated. For any given CV curve, its representation in the $\mathcal{GP}$ function space is obtained by finding a θ that maximizes the posterior probability $p(\mathbf{X} = I|\mathbf{y} = (v, t))$ [2]. We choose the $\mathcal{GP}$ model with a *non-stationary* covariance as a target model $\mathcal{M}_2$. It follows from the reproducing kernel Hilbert space (RKHS) theorem (Ch 12.4 in [5]) that any smooth function can be represented using a kernel or a covariance function. Thus for the null model ($\mathcal{M}_1$), it is sufficient to use a $\mathcal{GP}$ with smoothness controllable covariance function. A brief overview of the covariance functions selected as basis functions is described below. For both models $\mathcal{M}_1$ and $\mathcal{M}_2$, the mean function is chosen to be $\mu(x) = 0$ as we normalize the response curves $I(v, t)$ to be with in $(0, 1)$ and expect the covariance function to determine the shape of the CV curve.

[1] for this reason we use $k(x, x')$ and basis function of $\mathcal{GP}$ interchangeably in this paper

[2] we use the *maximum a posteriori* or MAP estimation

Squared Exponential Covariance

We use the commonly known *squared exponential kernel* (in Equation (5) and Figure 2) as a covariance model for $\mathcal{M}_1$ where the resulting feature map $\phi(x)$ forms a basis for functions that are smooth and stationary.

$$k(x, x') = \sigma_f^2 \exp\big((x - x')^\top \Lambda^{-1} (x - x')\big) \tag{5}$$

In Equation (5), σ_f is scaling parameter, and Λ is a diagonal matrix with each entry as a length scale for the corresponding dimension of $x, x' \in \mathcal{X}$. The left panel of Figure 2 depicts five samples drawn at random from the $\mathcal{GP}$ with the covariance in Equation (5).The right panel of the same figure depicts the covariance function visualized on a uniform grid of $\mathcal{X} \times \mathcal{X}$ as contours. From Figure 2, it can be seen that the covariance is stronger (≈ 1) between inputs with Euclidean norm (i.e. distance) less than a length scale controlled by the parameter Λ.

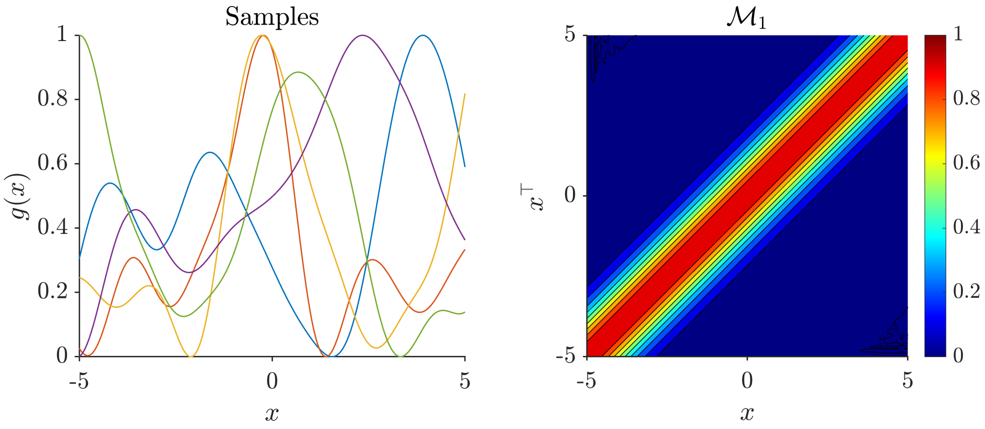

Figure 2: A pictorial representation of Equation (5). Left panel: five samples drawn at random from the $\mathcal{GP}$ built using Equation (5), captures the smooth and locally correlated nature of the $\mathcal{GP}$. Right panel: a contour plot depicting correlations between outputs of one-dimensional vectors $x, x' \in \mathcal{X}$. Color code represents the covariance $k(x, x')$ with red representing high covariance i.e. output values $g(x), g(x')$ are highly correlated and vice-versa.

Neural Network Covariance

We use a neural network covariance kernel to build a $\mathcal{GP}$ function space representation for the target model $\mathcal{M}_2$(shown in Equation (6) and Figure 3). The fast kinetic (or S-shape curve) responses have a non-stationary covariance and hence we choose a covariance that is effective in handling rapidly changing signals.

$$
\begin{aligned}
k(x, x') &= \sigma_f^2 \sin^{-1}\left(\frac{x^\top \Lambda^{-2} x'}{\sqrt{h(x)h(x')}}\right) \\
h(x) &= 1 + x^\top \Lambda^{-2} x
\end{aligned}
\tag{6}
$$

In Equation (6), σ_f is scaling parameter, and Λ is a diagonal matrix with each entry as a length scale. Figure 3 is analogues to Figure 2 and it can be seen that the covariance is high (≈ 1) in two blocks of input locations that are separated by a completely un-related input locations (covariance ≈ 0). This is in contrast to $\mathcal{M}_1$ where the covariance is determined by some form of distance between input points.

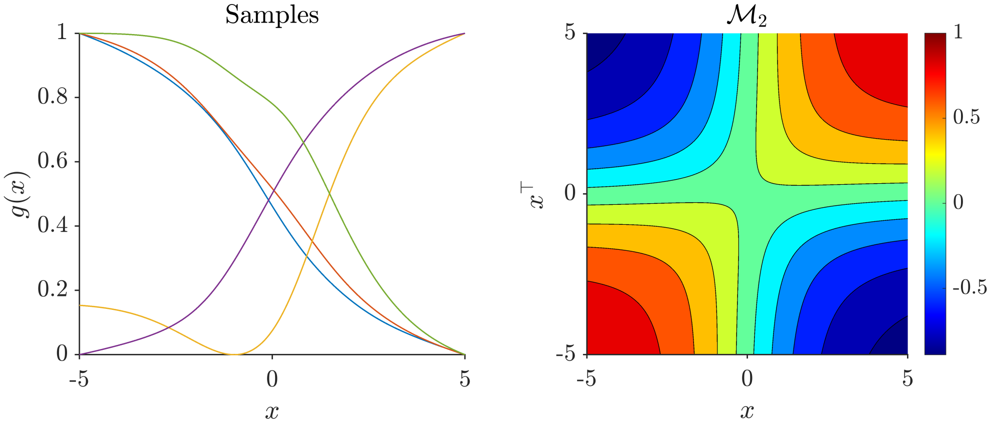

Figure 3: A pictorial representation of Equation (6). Left panel: five samples drawn at random from the $\mathcal{GP}$ built using Equation (6), captures the non-stationary nature nature of the $\mathcal{GP}$ signified by constant values and sharp rises in the response values. Right panel: a contour plot of covariance between two one-dimensional vectors $x, x' \in \mathbf{X}$ as inputs. A positive value for $k(x, x')$ signifies that output values $g(x), g(x')$ are highly correlated and vice-versa.

Results

To demonstrate the application of active search for S-shaped CV curves, we choose a classic EC-mechanism, that consists of two reactions: E and C (Equation (R1) and Equation (R2)) corresponding to one electron transfer reaction (E) and one chemical reaction (C), respectively. EC-mechanism is selected as it is a well studied mechanism [21, 3, 14] that produces a variety of CV shapes thus serves as a good test case for the oracle proposed in this paper. In this work, we use the MECSim [13] simulator to generate CV curves on demand.

Data generation

The EC mechanism is a two step reaction comprising of an electron transfer Equation (R1) followed by a chemical reaction in Equation (R2).

$$\mathrm{P} + \mathrm{e} \rightleftharpoons \mathrm{Q} \tag{R1}$$

$$\mathrm{Q} + \mathrm{A} \longrightarrow \mathrm{P} \tag{R2}$$

Electro-chemical kinetics of the EC mechanism can be modeled and solved using governing partial differential equations [22]. In this work, we are interested in modeling the kinetics of species (and electron) that contributes to current generation under cyclic voltage sweep at a given sweeping rate.

Towards this goal, the transport of the three species (P, Q, A) in the solution is modelled using Fick's second law of diffusion with a source term corresponding to the heterogeneous reactions:

$$\begin{aligned}
\frac{\partial C_P}{\partial t} &= D_{\text{diff}} \frac{\partial^2 C_P}{\partial u^2} + k_s C_Q C_A \\
\frac{\partial C_Q}{\partial t} &= D_{\text{diff}} \frac{\partial^2 C_Q}{\partial u^2} - k_s C_Q C_A \\
\frac{\partial C_A}{\partial t} &= D_{\text{diff}} \frac{\partial^2 C_A}{\partial u^2} - k_s C_Q C_A
\end{aligned} \tag{7}$$

with the boundary conditions defined as follows:

$$\begin{aligned}
&t = 0, \forall u && C_P = C_P^0, C_A = C_A^0, C_Q = C_Q^0 \\
&t > 0, u \to \infty && C_P = C_P^0, C_A = C_A^0, C_Q = C_Q^0 \\
&t > 0, \forall u && \frac{\partial C_A}{\partial u} = 0; \frac{\partial C_P}{\partial u} + \frac{\partial C_Q}{\partial u} = 0; C_P/C_Q = \exp\left(\frac{F}{RT}(V - E^0)\right)
\end{aligned} \tag{8}$$

In Equations (7) and (8) the formal reversible potential of electron transfer reaction eq. (R1) is E^0, the concentration of catalyst P is C_P, specie Q is C_Q and substrate A is C_A. D_{diff} is a common diffusion coefficient for all species and k_s is the rate constant of the forward reaction in Equation (R2). The spatial domain is denoted as u starting from the working electrode (i.e. $u = 0$) assuming a semi-infinite domain. The time scale of the simulation is denoted as t. Initial concentrations (i.e. at $t = 0$) are denoted with a superscript 0. V represents the time varying applied voltage. For a cyclic voltage sweep between voltages V_i, V_f at a rate of ν V/s we get Equation (9) for V (T_s is switching time).

$$V(t) = \begin{cases} V_i + \nu t & 0 < t < T_s \\ V_f - \nu t & T < t < 2T_s \end{cases} \tag{9}$$

Digital simulation of system of partial differential equations in Equations (7) and (8) is performed to determine spatio-temporal concentration profiles of species P, Q, A. The Faradaic current observed during the cyclic voltage load is computed using Equation (10) following the Butler-Volmer model for heterogeneous electron transfer at the electrode surface.

$$i(t, v) = F A_{\text{surf}} k^0 \left[C_Q \exp\left(\frac{\alpha F}{RT}(V - E^0)\right) - C_P \exp\left(\frac{(1-\alpha)F}{RT}(V - E^0)\right)\right] \tag{10}$$

In Equation (10), F is Faraday's constant, A_{surf} is surface area of electrode ($= 1$ cm^2), R is universal gas constant, T is room temperature. k^0 is heterogeneous electron transfer rate constant and α is a symmetric charge transfer coefficient (=0.5).

We use the freeware software MECSim [12, 13] to digitally simulate the cyclic voltammetry response in the voltage range of $[-0.5V, 0.5V]$ [3]. Along with the parameters used in Equations (7) and (8), MECSim can also simulate the effects of an uncompensated resistance (R_u), double layer capacitance (C_{dl}) which are not used in this work. We form a 6-dimensional design search space using $C_P^0, C_0^A, k_s, k^0, \nu, E^0$ and set the values of $R_u = 0, C_{dl} = 0, \alpha = 0.5, D = 1 \times 10^{-5}$. Table 1 lists the combinatorial space defined with six design variables (dimensions of search space) and number of samples along the dimension used to create an exhaustive search grid of tunable settings. After excluding responses from a diverging simulation arising from a combination of non-physical parameters for MECSim [4] we get a total of $\approx 17 \times 10^3$ CV curves in our database.

Table 1: Combinatorial space used to generate CV responses in EC mechanism along with number of levels used in the exhastive search.

Parameter	range	number of levels per dimension
$\log C_P^0$	[-2,3]	5
$\log C_A^0$	[-2,3]	5
E^0	[-0.4,0.4]	5
$\log k_s$	[-1,6]	5
$\log k^0$	[-1,6]	5
$\log \nu$	[-2,4]	6

Using BMS as an oracle to identify S-shaped CV curves

We demonstrate the application of the BMS oracle to label the CV responses as target, if they are of S-shape, and as non-targets otherwise. We use the proposed BMS oracle to label the CV responses and couple it with standard active search techniques to find our "targets" within a given budget of label queries (i.e. number of queries to the simulator). To accommodate for the high-throughput search running a batch of experiments at a time, we run the active search using both sequential selection of query locations (batch size $b = 1$) and a batch selection ($b = 100$). We use the design space in Table 1 and aim to find as many

[3] http://www.garethkennedy.net/MECSimDownload.html

[4] see MECSim documentation for known limitations

targets as possible in the resulting combinatorial design space $\mathcal{S}$ of six parameters (dimension of $\mathcal{S}$).

To demonstrate the efficacy of the proposed methods, we first pre-compute labels for set of ten CV curves in $\mathcal{S}$ of varying shape. Figure 4 depicts the chosen CV curves ordered based on model posterior (Equation (4)) percentile rank. Notably, the highest scored CV curves have the S-shape which are of interest in this work. From Figure 4, it can be noted that BMS assigned the highest score to CV curves where the forward and backward sweeps overlap exactly i.e. no hysteresis or capacitive behavior (highlighted using a red box). On the other spectrum, the oracle labels several types of CV curves with low scores. These types include the classic "duck-shape" curves, or curves that are diffusion driven with peaks in the forward and backward sweep.

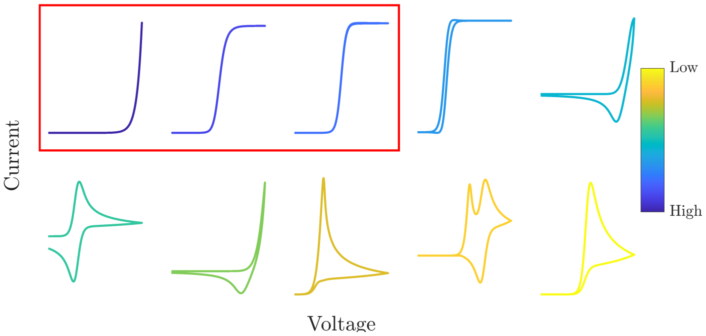

Figure 4: Representative CV curves from the dataset ordered and color coded using the BMS score. CV curves boxed in red will be labelled as targets by the oracle.

Active (Batch)Search for S-shaped CV curves

Active search with batch selection of locations in the design space has been recently studied and successfully applied to high throughput combinatorial search of material and drug discovery [10]. We use the state-of-the-art active batch search introduced in Jiang et.al [10], with a fixed budget of 1000 queries ($\approx 6\%$ of exhaustive search with details in Table 1) to the simulator for batch sizes of $b \in \{1, 100\}$ to actively query our combinatorial search space $\mathcal{S}$. For batch $b = 1$, the decision model $p(y = +1|\mathcal{D})$ is updated after each iteration, while for batch size $b = 100$, the decision model is updated after 100 CV measurements from the simulator and oracle. The batch size reflects the setting of the high throughput analysis, as often material is prepared in batches.

A label for any given location is assigned based on the application of BMS oracle to the corresponding CV curve $I(v,t)$ simulated by solving Equations (7) and (8) over 4×10^3 discrete time points. We label a CV response as target if its BMS oracle score is in the range defined by top three percentile ranks [5] [6] shown in Figure 4.

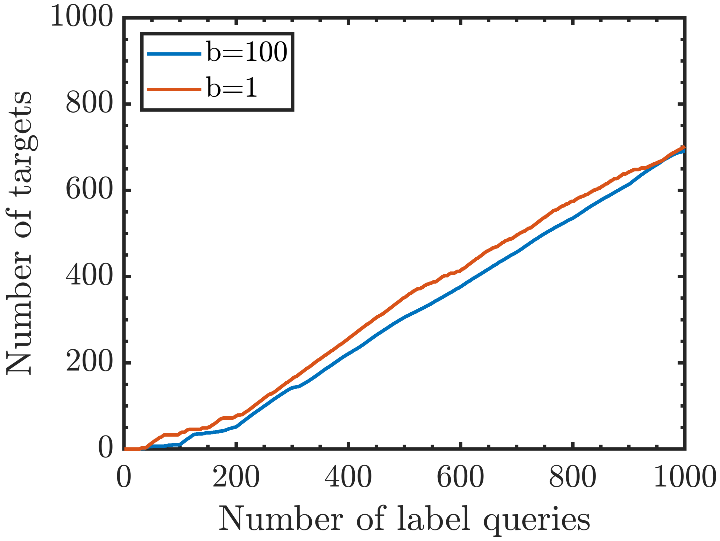

Figure 5: Active target detection in the EC mechanism combinatorial search space (see Table 1 for definition of search space). We repeat the active search 20 times, each time starting with a randomly chosen non-S-shape data point in $\mathcal{S}$.

We assume that our design space is continuous thus a k-nearest neighbor probability distribution is used a decision model in Bayesian active learning following the approach in [10]. This assumption implies that if we find a target at a certain location in the search space, k-closest neighbors in the design space also are highly likely to be a target as well.

In Figure 5, we report the average number of targets found in the design space over the number of label queries for two batch sizes ($b = 1, 100$) considered [7]. Our results demonstrate

[5] this is a heuristic and can be altered based on application

[6] Similarly for active search of bi-functional oxygen electrocatalysts, one can assign a material as a target if both of its OER and ORR experimental CV curves are in the top three percentile ranks of BMS scores.

[7] The number of target are averaged over a total of 20 active searches each time start with a randomly

that searching the design space using active learning can be useful, with a near linear target detection. It can also be noted from Figure 5 for any given number of allowed label queries to the oracle (or equivalently number of simulation queries to the simulator), the sequential selection finds marginally more targets than the batch selection $b = 100$. This observation is in accordance with Theorem 1 in [10]. Jiang et.al, [10] argue that batch selection suffers from having to select a batch from the search space with fewer observed responses and locations. However, from experimental point of view, one need to consider the advantages and dis-advantages of sequential selection over batch selection.

Conclusion and future work

In conclusion, we defined and evaluated a $\mathcal{GP}$-based oracle for materials discovery using cyclic voltammetry. Next, we combined the oracle with a state-of-the-art active batch search to identify condition resulting in the targeted shape of CV curve. We demonstrated a robust high throughput combinatorial search to find the target responses using only $< 6\%$ of total number of CV experiments from the corresponding exhaustive search (with a discrete sampling of modest 5 levels per dimension).

This work has implications in identification of characterization conditions where kinetic knowledge extraction from the cyclic voltammetry can be preformed more effectively. Specifically, we have illustrated a framework that can be used to identify S-shaped CV curves. Once S-shaped CV curve is obtained, a foot of the wave analysis can be applied [25] to extract rate constant for rate determining step, overpotential dependent turn over frequency etc. In this sense our method has applications in accelerated knowledge extraction, with the application in screening for target catalysts including the bi-functional alkaline fuel cell catalysts that motivated this work.

selected sample in the search space

Acknowledgement

The authors thank Toyota Research Institute for providing the support through Accelerated Material Design Discovery program.

Code and data availability

All the data and code to reproduce the experiments from this paper can be found at `https://github.com/kiranvad/gpcv`

References

[1] Allen J Bard, Larry R Faulkner, et al. “Fundamentals and applications”. In: *Electrochemical Methods* 2.482 (2001), pp. 580–632.

[2] Kieren Bradley et al. “Reversible perovskite electrocatalysts for oxygen reduction/oxygen evolution”. In: *Chemical science* 10.17 (2019), pp. 4609–4617.

[3] Cyrille Costentin and Jean-Michel Saveant. “Cyclic voltammetry analysis of electrocatalytic films”. In: *The Journal of Physical Chemistry C* 119.22 (2015), pp. 12174–12182.

[4] Cyrille Costentin et al. “Turnover numbers, turnover frequencies, and overpotential in molecular catalysis of electrochemical reactions. Cyclic voltammetry and preparative-scale electrolysis”. In: *Journal of the American Chemical Society* 134.27 (2012), pp. 11235–11242.

[5] Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.

[6] Jacob R Gardner et al. “Psychophysical Detection Testing with Bayesian Active Learning.” In: *UAI*. 2015, pp. 286–295.

[7] Roman Garnett et al. “Bayesian optimal active search and surveying”. In: *arXiv preprint arXiv:1206.6406* (2012).

[8] David J Gavaghan et al. “Use of Bayesian inference for parameter recovery in DC and AC Voltammetry”. In: *ChemElectroChem* 5.6 (2018), pp. 917–935.

[9] Joel A Haber et al. “High-Throughput Mapping of the Electrochemical Properties of (Ni-Fe-Co-Ce) Ox Oxygen-Evolution Catalysts”. In: *ChemElectroChem* 1.3 (2014), pp. 524–528.

[10] Shali Jiang et al. “Efficient nonmyopic active search with applications in drug and materials discovery”. In: *arXiv preprint arXiv:1811.08871* (2018).

[11] Jae-Il Jung et al. “Optimizing nanoparticle perovskite for bifunctional oxygen electrocatalysis”. In: *Energy & Environmental Science* 9.1 (2016), pp. 176–183.

[12] Gareth Kennedy. *Monash Electrochemistry Simulator (MECSim)*. 2015.

[13] Gareth F Kennedy, Alan M Bond, and Alexandr N Simonov. “Modelling ac voltammetry with MECSim: facilitating simulation–experiment comparisons”. In: *Current Opinion in Electrochemistry* 1.1 (2017), pp. 140–147.

[14] Gareth F Kennedy, Jie Zhang, and Alan M Bond. “Automatically identifying electrode reaction mechanisms using deep neural networks”. In: *Analytical chemistry* 91.19 (2019), pp. 12220–12227.

[15] Jiezhen Li et al. “Application of Bayesian Inference in Fourier-Transformed Alternating Current Voltammetry for Electrode Kinetic Mechanism Distinction”. In: *Analytical chemistry* 91.8 (2019), pp. 5303–5309.

[16] Daniel J Martin et al. “Qualitative extension of the EC Zone Diagram to a molecular catalyst for a multi-electron, multi-substrate electrochemical reaction”. In: *Dalton Transactions* 45.24 (2016), pp. 9970–9976.

[17] Roc Matheu et al. “Foot of the wave analysis for mechanistic elucidation and benchmarking applications in molecular water oxidation catalysis”. In: *ChemSusChem Communications* (2016).

[18] Jens Kehlet Nørskov et al. “Origin of the overpotential for oxygen reduction at a fuel-cell cathode”. In: *The Journal of Physical Chemistry B* 108.46 (2004), pp. 17886–17892.

[19] Dino Oglic et al. “Active search for computer-aided drug design”. In: *Molecular informatics* 37.1-2 (2018), p. 1700130.

[20] Martin Robinson et al. “Separating the Effects of Experimental Noise from Inherent System Variability in Voltammetry: The [Fe (CN)] ^{3}–/–Process”. In: (2018).

[21] Eric S Rountree et al. *Evaluation of homogeneous electrocatalysts by cyclic voltammetry.* 2014.

[22] JM Savéant and KB Su. “Homogeneous redox catalysis of electrochemical reaction: Part VI. Zone diagram representation of the kinetic regimes”. In: *Journal of electroanalytical chemistry and interfacial electrochemistry* 171.1-2 (1984), pp. 341–349.

[23] Helge S Stein et al. “Functional mapping reveals mechanistic clusters for OER catalysis across (Cu–Mn–Ta–Co–Sn–Fe) O x composition and pH space”. In: *Materials Horizons* (2019).

[24] Santosh K Suram et al. “Generating information-rich high-throughput experimental materials genomes using functional clustering via multitree genetic programming and information theory”. In: *ACS combinatorial science* 17.4 (2015), pp. 224–233.

[25] Vincent C-C Wang and Ben A Johnson. “Interpreting the Electrocatalytic Voltammetry of Homogeneous Catalysts by the Foot of the Wave Analysis and Its Wider Implications”. In: *ACS Catalysis* 9.8 (2019), pp. 7109–7123.

[26] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning.* Vol. 2. 3. MIT Press Cambridge, MA, 2006.