ENM360 Final Paper: Applications of Gaussian Process Regression In the context of Optimizing The Design of Organic Photovoltaic Devices.

Written by
Rachael Tamakloe

-Abstract:

In one of the latter sections of this course, we were introduced to optimizing sequential decision making through the use of Baysesian optimization. In the applied context of materials design, the sequential decision making process focuses on iteratively choosing design parameters that aim to maximize the performance of a given material design. The problem setting revolves around the goal of recommending promising design candidates in pursuit of optimizing some measure of quality. Given some recorded data, the first goal is to learn a function that relates the design variables to target quantities of interest, and the second goal is to recommend promising design candidates to test in pursuit of finding an optimal design with the best performance (PredictiveIntelligenceLab, 2021).

Building on the knowledge gained from this course, my project focuses on a two-fold objective. First, I will cover the use of Gaussian Process Regression as a way to learn an unknown function of interest, and second, I will apply the use of Gaussian Process regression in the context of optimizing the design of organic photovoltaic devices. I will begin by briefly introducing the motivations for the need of Gaussian Process Regression in complex experimental settings and then transition to a focus on its application in a real world example of optimizing the performance of organic photovoltaic devices.

-Introduction:

In the context of materials design, optimal designs are achieved through the traditional approach of the iterative process of trial and error through repeated experimentation. This trial and error process first begins when some material designs are first constructed (based on knowledge and intuition from past experience), and then tested in experiments to see how well they perform. Using what is then learned from these experiments, other designs are chosen next to implement and test during experimentation (Frazier & Wang, 2015). This process is generally repeated until some acceptable/optimal design has been achieved.

In academic laboratories one factor/one variable at a time (OFAT, or OVAT) is an approach to experimentation in which one variable is changed at a time in order not to confound the roles of different factors that lead to an observed result. In large multi-dimensional material design systems, this approach is slow and infrequently results in the discovery of optimal solutions (Adutwum et al., 2018). Additionally, there is no confidence in knowing whether one’s result is truly the most optimal one, since a large space of values/parameters of the system would not have been tested (Adutwum et al., 2018).

In many cases where repetitive experimentation is either time consuming or expensive, there arrives the natural need for a more efficient strategy/technique that is able to accelerate the achievement of an optimal design through the trial and error process. One proposed technique to solve this need is the use of Bayesian optimization in Gaussian Process Regression to aid in deciding which designs are more valuable to try in the next iteration.

In contrast to the traditional approach, the application of Bayesian Optimization can significantly increase the rate of screening and optimization of different material properties. With the use of Bayesian Optimization in mind, experiments are first carried out to sample a large, multidimensional parameter space in an efficient manner (Adutwum et al., 2018). Using the collected data, a parameter space can then be mapped, enabling the ability to explore new parameters, with the goal of finding design parameters in the parameter space that optimize a material design system.

Because most well-developed Bayesian optimization methods assume a vector of continuous quantitative variables as inputs, this subset of Bayesian optimization will be the focus of my project specifically in the context of materials design.

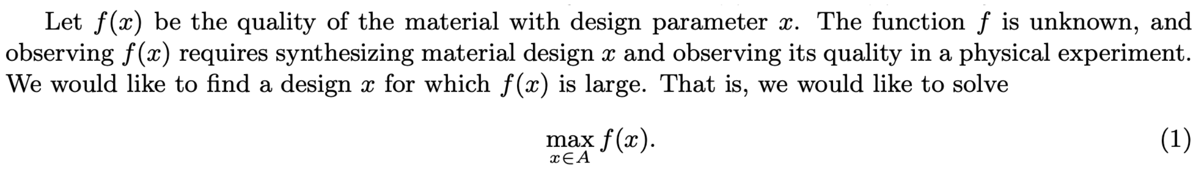
The Bayesian Optimization method of Gaussian Process Regression will be well suited for cases where (Frazier & Wang, 2015):

1. Input parameters are a vector of quantitative continuous variables
2. There’s only a single measure of quality that we aim to make as large as possible
3. The constraints on possible materials designs are all incorporated into the quality measure.

-Formal Mathematical Definition of the Problem

Before I transition to the application of Gaussian Process Regression in optimizing organic photovoltaic devices, I will first present the mathematical model behind the Gaussian Process Regression model.

Given some recorded data, the goal of Gaussian Process Regression is to learn a function that links the design variables to target quantities of interest. In the case of material design, this target quantity will generally be some measure of quality, *f(x)*, parametrized by its inputs, *x*.

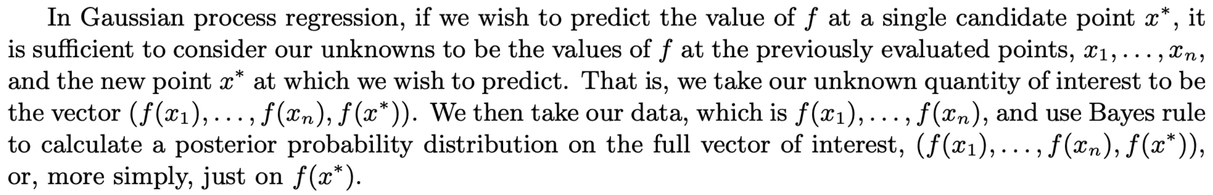


Let $f(x)$ be the quality of the material with design parameter x. The function f is unknown, and observing $f(x)$ requires synthesizing material design x and observing its quality in a physical experiment. We would like to find a design x for which $f(x)$ is large. That is, we would like to solve

$$\max_{x \in A} f(x). \tag{1}$$

(Frazier & Wang, 2015)

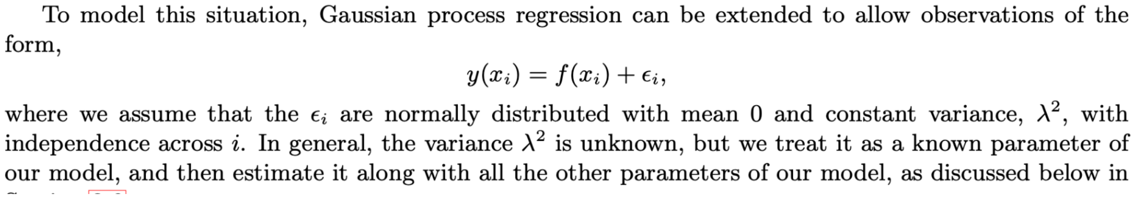
Gaussian process regression estimates f(x) by placing a multivariate gaussian prior probability distribution on unknown quantities of interest. Using Bayes rule together with previously recorded data observed, a posterior probability distribution is calculated on these unknowns.



In Gaussian process regression, if we wish to predict the value of f at a single candidate point x^*, it is sufficient to consider our unknowns to be the values of f at the previously evaluated points, $x_1, \ldots, x_n$, and the new point x^* at which we wish to predict. That is, we take our unknown quantity of interest to be the vector $(f(x_1), \ldots, f(x_n), f(x^*))$. We then take our data, which is $f(x_1), \ldots, f(x_n)$, and use Bayes rule to calculate a posterior probability distribution on the full vector of interest, $(f(x_1), \ldots, f(x_n), f(x^*))$, or, more simply, just on $f(x^*)$.

(Frazier & Wang, 2015)

In most real world physical experiments however, observations measured include noise.



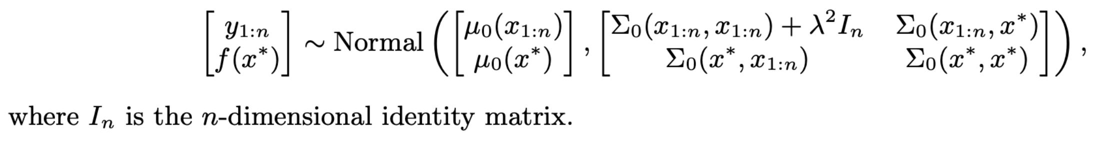
To model this situation, Gaussian process regression can be extended to allow observations of the form,

$$y(x_i) = f(x_i) + \epsilon_i,$$

where we assume that the ϵ_i are normally distributed with mean 0 and constant variance, λ^2, with independence across i. In general, the variance λ^2 is unknown, but we treat it as a known parameter of our model, and then estimate it along with all the other parameters of our model, as discussed below in

(Frazier & Wang, 2015)

Accounting for noisy observation in our recorded data, our prior then becomes:

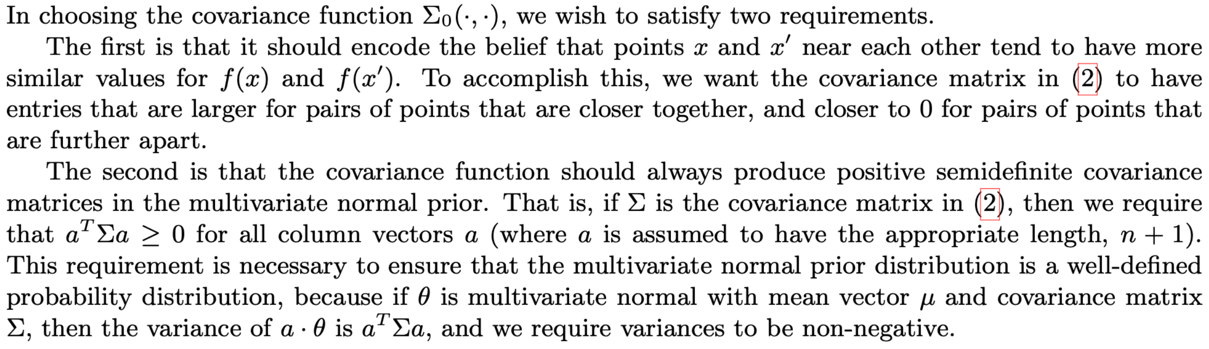


$$\begin{bmatrix} y_{1:n} \\ f(x^*) \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} \mu_0(x_{1:n}) \\ \mu_0(x^*) \end{bmatrix}, \begin{bmatrix} \Sigma_0(x_{1:n}, x_{1:n}) + \lambda^2 I_n & \Sigma_0(x_{1:n}, x^*) \\ \Sigma_0(x^*, x_{1:n}) & \Sigma_0(x^*, x^*) \end{bmatrix} \right),$$

where I_n is the n-dimensional identity matrix.

(Frazier & Wang, 2015)

To generate the covariance matrix of the multivariate normal prior distribution, a covariance function or covariance kernel is used. Shown above as Σ0(·,·), the covariance function takes a pair of points x,x′ as inputs. It applies this covariance function to every pair of points in x1,...,xn, and x* to create an (n+1)×(n+1) matrix.



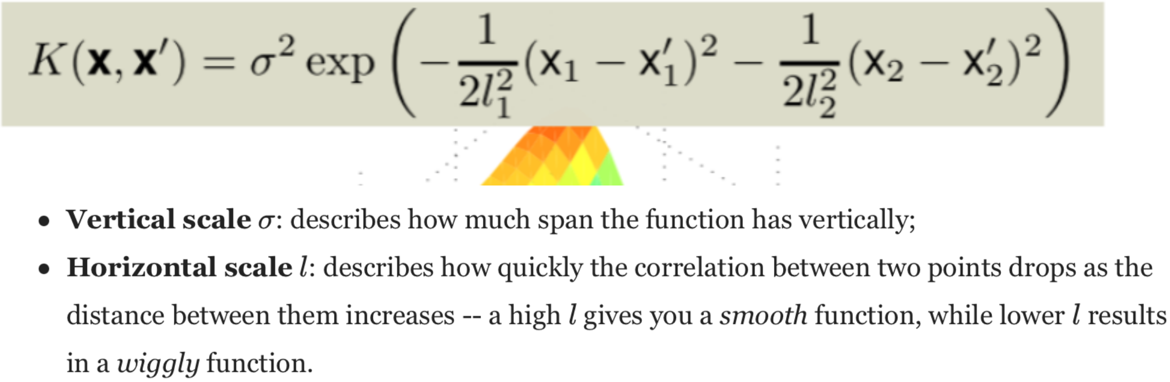
In choosing the covariance function $\Sigma_0(\cdot,\cdot)$, we wish to satisfy two requirements.

The first is that it should encode the belief that points x and x' near each other tend to have more similar values for $f(x)$ and $f(x')$. To accomplish this, we want the covariance matrix in (2) to have entries that are larger for pairs of points that are closer together, and closer to 0 for pairs of points that are further apart.

The second is that the covariance function should always produce positive semidefinite covariance matrices in the multivariate normal prior. That is, if Σ is the covariance matrix in (2), then we require that $a^T\Sigma a \geq 0$ for all column vectors a (where a is assumed to have the appropriate length, $n+1$). This requirement is necessary to ensure that the multivariate normal prior distribution is a well-defined probability distribution, because if θ is multivariate normal with mean vector μ and covariance matrix Σ, then the variance of $a \cdot \theta$ is $a^T\Sigma a$, and we require variances to be non-negative.

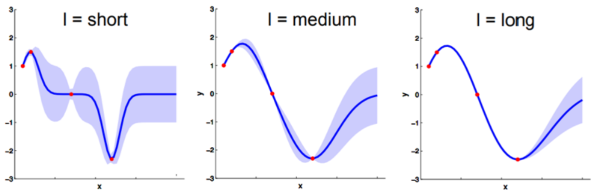
(Frazier & Wang, 2015)

An example of a very common covariance function is the RBF Kernel. Pictured below in the case of 2-dimensional input data, the RBF kernel has d+1 parameters, where the first parameter sigma, **σ**, encodes how much variability there is present in vertical span of the function, and the rest of the d parameters encode how much correlation there is between two points at a given dimension.



$$K(\mathbf{x},\mathbf{x}') = \sigma^2 \exp\left(-\frac{1}{2l_1^2}(\mathsf{x}_1 - \mathsf{x}_1')^2 - \frac{1}{2l_2^2}(\mathsf{x}_2 - \mathsf{x}_2')^2\right)$$

- **Vertical scale** σ: describes how much span the function has vertically;
- **Horizontal scale** l: describes how quickly the correlation between two points drops as the distance between them increases -- a high l gives you a *smooth* function, while lower l results in a *wiggly* function.



(Shi, 2021)

In order to find the right values for these d+1 parameters of the covariance function pictured above, empirical methods such as MLE can be used to estimate them by maximizing a log marginal likelihood with the aim of finding the hyperparameters that make the observed data as likely as possible.

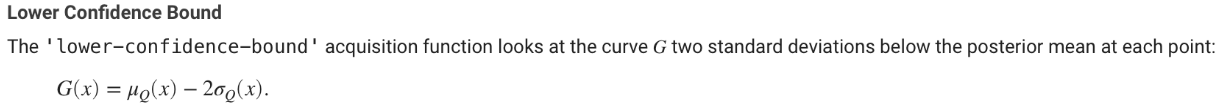
Using qualities about the conditional probabilities of multivariate normal distributions, it can be arrived that the posterior distribution on f(x∗) given observations yi = f(xi), i = 1,..., n is normal, with a mean μn(x*) and variance σn^2 (x*) is given by:

$$\begin{aligned}\mu_n(x^*) &= \mu_0(x^*) + \Sigma_0(x^*, x_{1:n}) \left[\Sigma_0(x_{1:n}, x_{1:n}) + \lambda^2 I_n\right]^{-1} (y_{1:n} - \mu_0(x_{1:n})) \\ \sigma_n^2(x^*) &= \Sigma_0(x^*, x^*) - \Sigma_0(x^*, x_{1:n}) \left[\Sigma_0(x_{1:n}, x_{1:n}) + \lambda^2 I_n\right]^{-1} \Sigma_0(x_{1:n}, x^*).\end{aligned}$$

(Frazier & Wang, 2015)

We can think about our posterior mean, μn(x∗), as our estimate for f(x*), and our posterior variance, σn^2 (x∗), as a measure of uncertainty on our belief on f(x*).

Through the model of Gaussian Process regression described above, we have a way to learn our unknown function of interest. Naturally the next step is to develop a way to recommend new parameters to test. This can be done by optimizing over acquisition functions to find promising parameters to test for in new experiments. There are various acquisition functions that can be used. A common one is the lower confidence bound pictured below:



Lower Confidence Bound

The `'lower-confidence-bound'` acquisition function looks at the curve G two standard deviations below the posterior mean at each point:

$$G(x) = \mu_Q(x) - 2\sigma_Q(x).$$

(*Bayesopt*. Bayesian Optimization Algorithm)

The x which maximizes our acquisition function will be the recommendation for the next material design to implement and test.

-Application of Gaussian Process Regression in Optimizing the design of Organic Photovoltaics:

Organic Photovoltaics (OPV) are devices that convert solar energy to electrical energy. Most OPV devices consist of a layer of photoactive materials placed between two electrodes (Organic Photovoltaics - Sigmaaldrich.com). This layer of photoactive materials is called the bulk heterojunction (BHJ). The architecture and composition of the BHJ plays a very significant role in the performance of an OPV device. At least thousands of different BHJs have been tested and the performance of these devices depend greatly on the morphology of the BHJ that results from design parameters chosen (Adutwum et al., 2018). Because the options for potential components in a given OPV device is vast, it is impossible to test every combination due to a lack of time and resources. As a result, in a possible multi-dimensional material design system such as this, there is the natural need to use some machine learning technique that maps the parameter space to an estimated measure of quality (Adutwum et al., 2018). Having this estimation accelerates the iterative process of the search for the optimal design parameters. When evaluating the performance of an OPV device, a common measure of quality used is the Power of Conversion Efficiency, which is the ratio between the amount of useful power produced and the amount of total power absorbed/received (2021).

In this section of my paper, I will try to replicate the analytical results of a past organic photovoltaic optimization design experiment that was performed using support vector machines and an RBF kernel. The input parameters of this experiment were the donor weight percentage (wt %), the total solution concentration (mg/mL), and the bulk heterojunction spin-cast speed (rpm) of a given OPV design. The measure of quality used was the power conversion efficiency.

The experiment had two rounds of data collection. The first round consisted of using a latin square sampling technique to test out 16 different design parameter inputs. These 16 different designs were chosen in a manner that enabled exploration of the parameter space in order to approximate the functional dependence of PCE on the input parameters (Adutwum et al., 2018). Shown below is the range of the parameter space tested and the data gathered from the 16 different experiments:

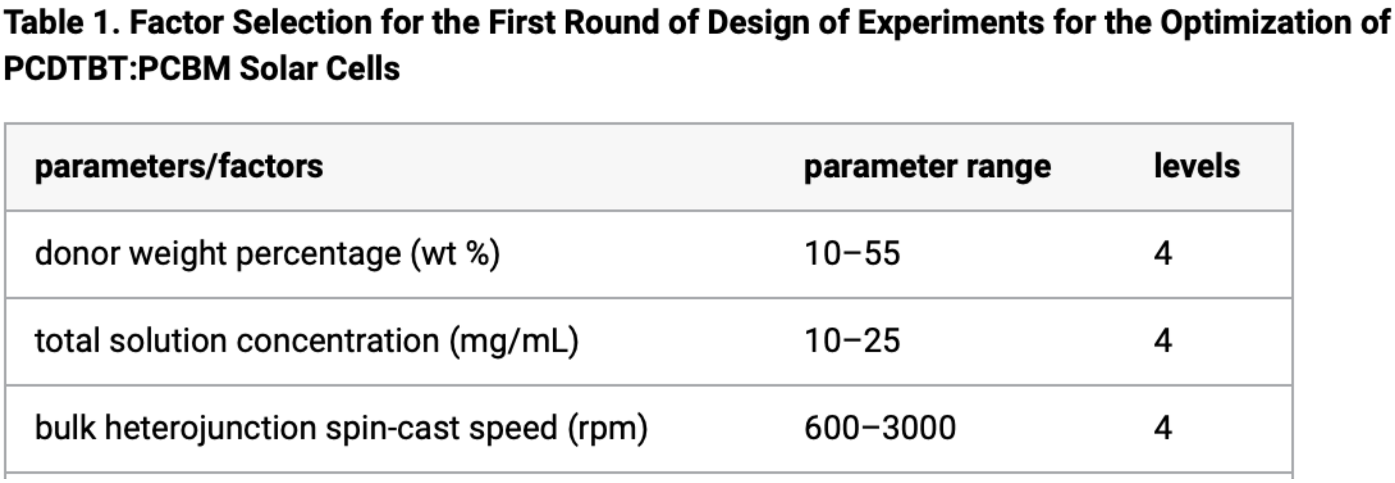
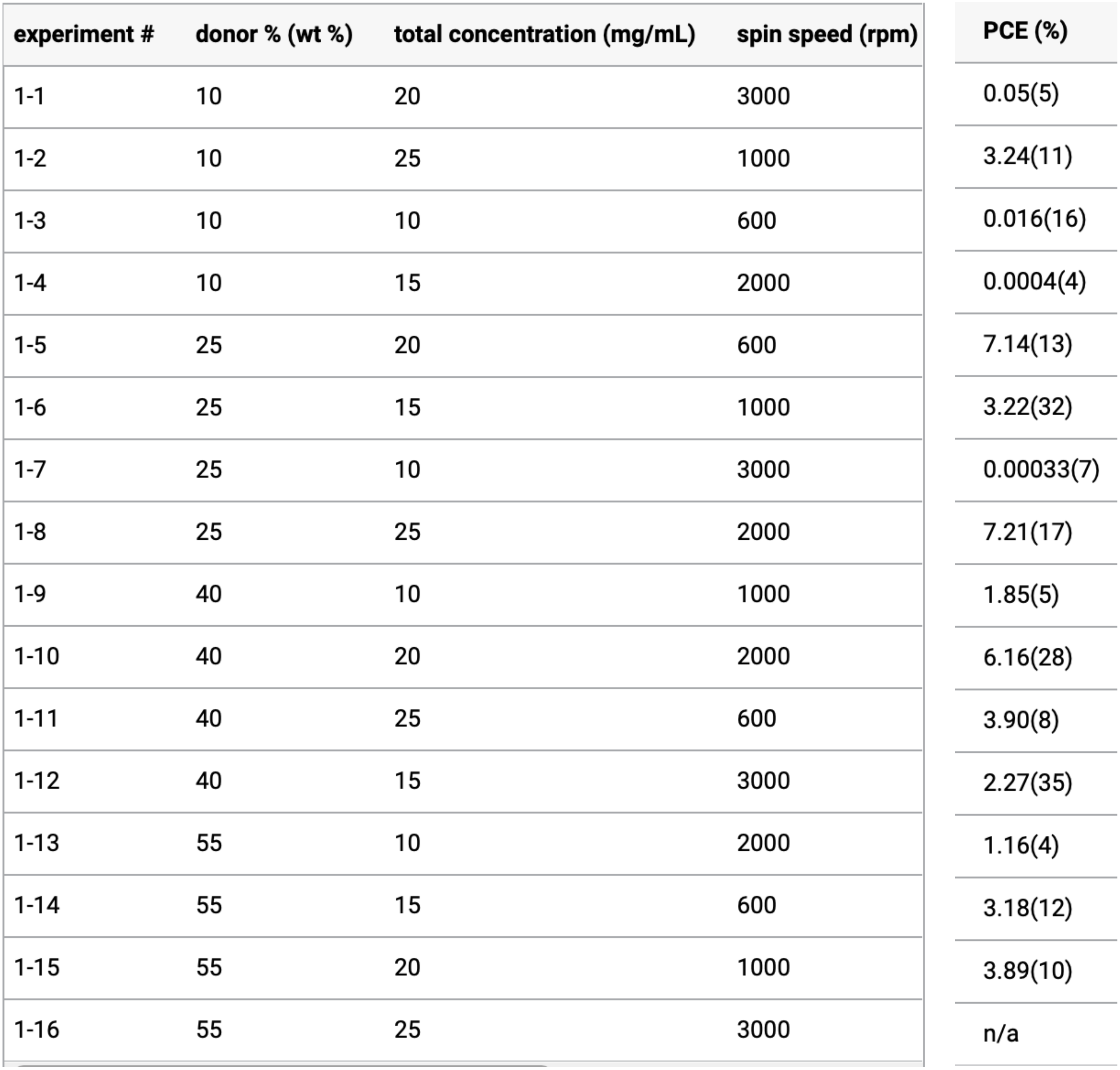


Table 1. Factor Selection for the First Round of Design of Experiments for the Optimization of PCDTBT:PCBM Solar Cells

parameters/factors	parameter range	levels
donor weight percentage (wt %)	10–55	4
total solution concentration (mg/mL)	10–25	4
bulk heterojunction spin-cast speed (rpm)	600–3000	4

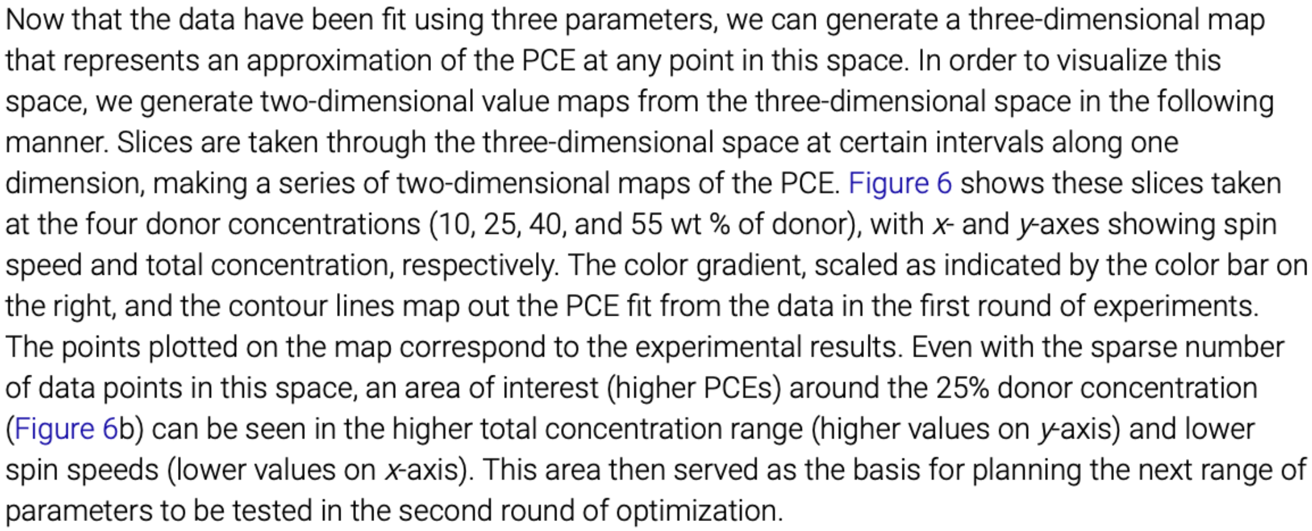
(Adutwum et al., 2018)



experiment #	donor % (wt %)	total concentration (mg/mL)	spin speed (rpm)	PCE (%)
1-1	10	20	3000	0.05(5)
1-2	10	25	1000	3.24(11)
1-3	10	10	600	0.016(16)
1-4	10	15	2000	0.0004(4)
1-5	25	20	600	7.14(13)
1-6	25	15	1000	3.22(32)
1-7	25	10	3000	0.00033(7)
1-8	25	25	2000	7.21(17)
1-9	40	10	1000	1.85(5)
1-10	40	20	2000	6.16(28)
1-11	40	25	600	3.90(8)
1-12	40	15	3000	2.27(35)
1-13	55	10	2000	1.16(4)
1-14	55	15	600	3.18(12)
1-15	55	20	1000	3.89(10)
1-16	55	25	3000	n/a

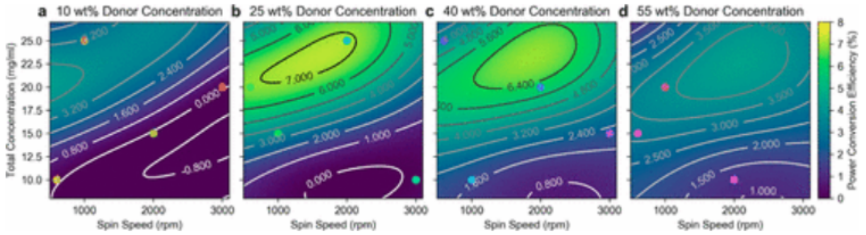
(Adutwum et al., 2018)

Using support vector machines and an RBF kernel, the 16 data points were then used to approximate the value of PCE at any point in the parameter space.



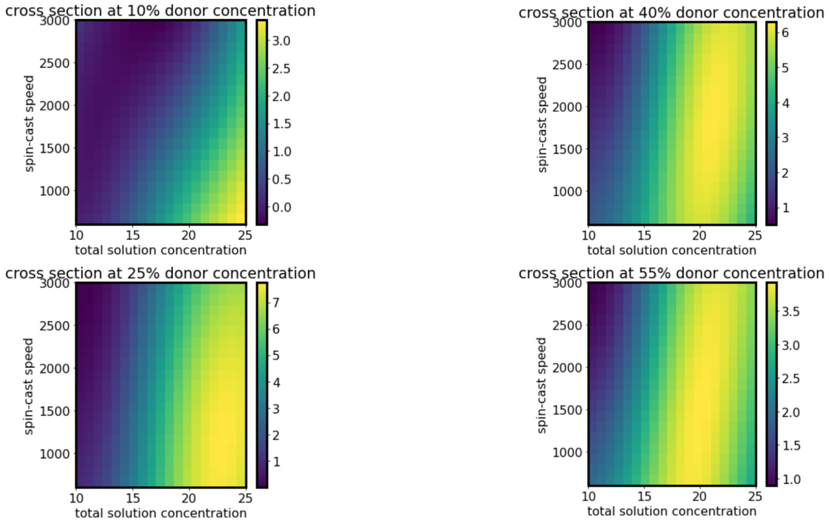
Now that the data have been fit using three parameters, we can generate a three-dimensional map that represents an approximation of the PCE at any point in this space. In order to visualize this space, we generate two-dimensional value maps from the three-dimensional space in the following manner. Slices are taken through the three-dimensional space at certain intervals along one dimension, making a series of two-dimensional maps of the PCE. Figure 6 shows these slices taken at the four donor concentrations (10, 25, 40, and 55 wt % of donor), with *x*- and *y*-axes showing spin speed and total concentration, respectively. The color gradient, scaled as indicated by the color bar on the right, and the contour lines map out the PCE fit from the data in the first round of experiments. The points plotted on the map correspond to the experimental results. Even with the sparse number of data points in this space, an area of interest (higher PCEs) around the 25% donor concentration (Figure 6b) can be seen in the higher total concentration range (higher values on *y*-axis) and lower spin speeds (lower values on *x*-axis). This area then served as the basis for planning the next range of parameters to be tested in the second round of optimization.

Figure 6



(Adutwum et al., 2018)

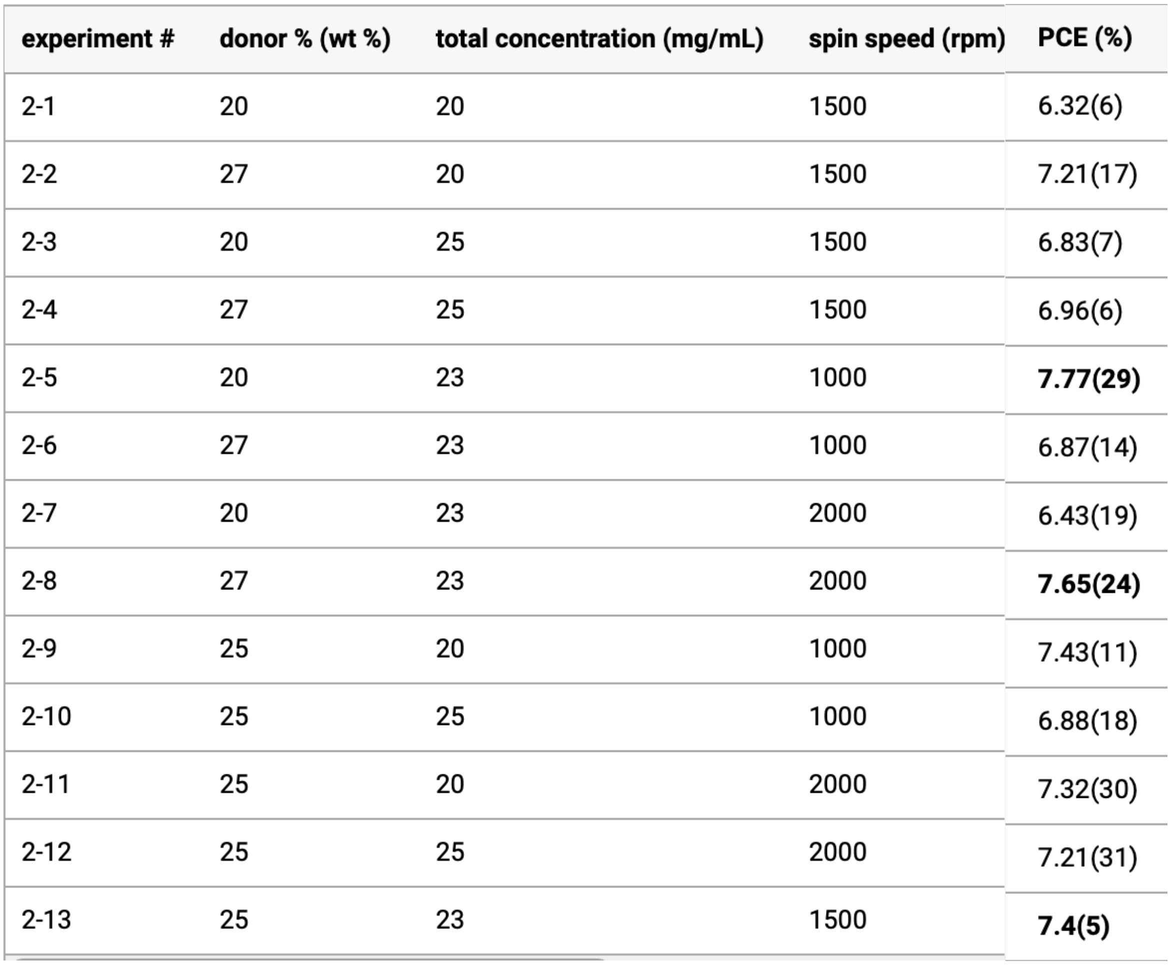
Using Gaussian process regression and an RBF kernel for the choice of covariance function, I was able to replicate similar analytical results using the observed data:



Analyzing cross sections of the parameter space at specific donor concentrations, we can make a few observations:

1) By taking a look at the maximum of the gradient bars for each of the 4 charts above, It is evident that an optimal donor concentration lives in the range from 25% to 40%.
2) An optimal total solution concentration lives in the range from 20 to 25(mg/mL).
3) An optimal spin cast speed lives in the range from 1000 to 2000(rpm).

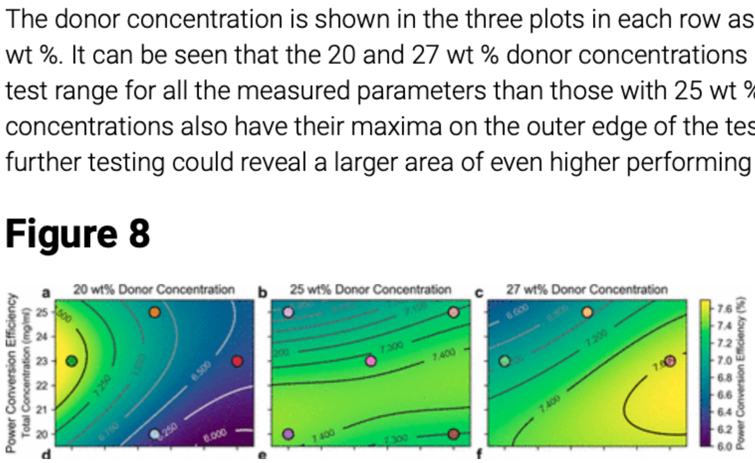
From these new insights, new inputs chosen to be tested can be picked from a narrower range in the parameter space for the second round of data collection. Pictured below is data gathered from the 13 different experiments done in the second round:



experiment #	donor % (wt %)	total concentration (mg/mL)	spin speed (rpm)	PCE (%)
2-1	20	20	1500	6.32(6)
2-2	27	20	1500	7.21(17)
2-3	20	25	1500	6.83(7)
2-4	27	25	1500	6.96(6)
2-5	20	23	1000	**7.77(29)**
2-6	27	23	1000	6.87(14)
2-7	20	23	2000	6.43(19)
2-8	27	23	2000	**7.65(24)**
2-9	25	20	1000	7.43(11)
2-10	25	25	1000	6.88(18)
2-11	25	20	2000	7.32(30)
2-12	25	25	2000	7.21(31)
2-13	25	23	1500	**7.4(5)**

(Adutwum et al., 2018)

Using support vector machines and an RBF kernel, the 13 new data points are added to the previous 15 data points and then are used to approximate the value of PCE at any point in the parameter space. Pictured below are the analytical results from the PCE estimation:



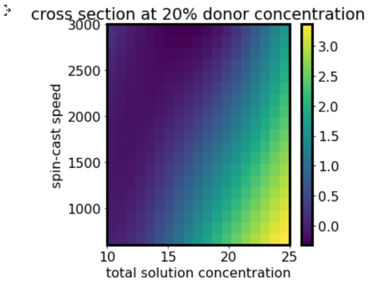
The donor concentration is shown in the three plots in each row as a slice of the RBF at 20, 25, and 27 wt %. It can be seen that the 20 and 27 wt % donor concentrations have more variability within the test range for all the measured parameters than those with 25 wt %. The 20 and 27 wt % donor concentrations also have their maxima on the outer edge of the test range. These results indicate that further testing could reveal a larger area of even higher performing OPV devices.

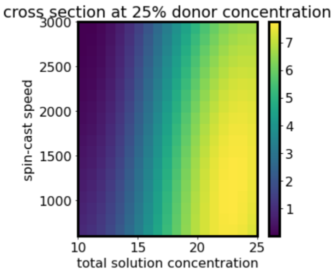
Figure 8

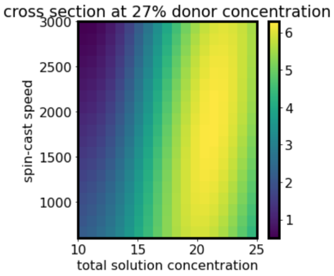
a
20 wt% Donor Concentration
b
25 wt% Donor Concentration
c
27 wt% Donor Concentration
Power Conversion Efficiency
Total Concentration (mg/ml)
25
24
23
22
21
20
7.250
6.750
6.500
6.250
6.000
7.300
7.400
6.600
7.200
d
e
f
7.6
7.4
7.2
7.0
6.8
6.6
6.4
6.2
6.0
Power Conversion Efficiency (%)

(Adutwum et al., 2018)

Using Gaussian process regression and an RBF kernel for the choice of covariance function, I was able to replicate similar analytical results using the observed data:







Analyzing cross sections of the parameter space at specific donor concentrations, we can make a few observations:

4) By taking a look at the range of the gradient bars for each of the 3 charts above, It is evident that there’s more variability present in PCE for the charts with 25% weight donor concentration and 27% weight donor concentration.

5) Comparing my charts produced by Gaussian process regression to the charts produced by SVM, I note a few differences in our PCE estimations in the case of the 20% weight donor concentration.
6) Overall, comparing the significant improvement of the average PCE in data collection from round 1 to round 2, we note the effectiveness of the use of Gaussian process regression as a way to accelerate the achievement of an optimal design.

The growing implementation of Bayesian Optimization in materials design gives way for more time efficient and possibly more money efficient ways for finding optimal material designs in real world applications where repeated trial and error experimentation can not produce effective results due to time/financial limitations. Furthermore, with the emergence of materials databases that make it easier to have access to existing data as a starting point for Bayesian optimization, the implementation of Bayesian optimization in materials design will definitely grow to become a new norm (Zhang et al., 2020).

References:

Adutwum, L. A., Olsen, B. C., Mar, A., & Buriak, J. M. (2018, July 20). *How to optimize materials and devices via design of experiments and Machine Learning: Demonstration Using Organic Photovoltaics*. ACS Publications. Retrieved December 22, 2021, from https://pubs.acs.org/doi/10.1021/acsnano.8b04726

Bayesopt. Bayesian Optimization Algorithm - MATLAB & Simulink. (n.d.). Retrieved December 23, 2021, from https://www.mathworks.com/help/stats/bayesian-optimization-algorithm.html#bva8re7-1

Frazier, P. I., & Wang, J. (2015). Bayesian Optimization for Materials Design. *Information Science for Materials Discovery and Design Springer Series in Materials Science,*45-75. doi:10.1007/978-3-319-23871-5_3

Organic Photovoltaics - Sigmaaldrich.com. (n.d.). Retrieved December 22, 2021, from https://www.sigmaaldrich.com/US/en/technical-documents/technical-article/materials-science-and-engineering/photovoltaics-and-solar-cells/opv-tutorial

PredictiveIntelligenceLab, P. P. (2021, December 2). *PredictiveIntelligenceLab/ENM360*. GitHub. Retrieved December 23, 2021, from https://github.com/PredictiveIntelligenceLab/ENM360

Shi, Y. (2021, December 5). *Gaussian processes, not quite for dummies*. The Gradient. Retrieved December 23, 2021, from https://thegradient.pub/gaussian-process-not-quite-for-dummies/

Wikimedia Foundation. (2021, October 21). *Energy conversion efficiency*. Wikipedia. Retrieved December 23, 2021, from https://en.wikipedia.org/wiki/Energy_conversion_efficiency

Zhang, Y., Apley, D. W., & Chen, W. (2020, March 18). Bayesian Optimization for Materials Design with Mixed Quantitative and Qualitative Variables. Retrieved from https://www.nature.com/articles/s41598-020-60652-9