Report on Project 2b: Bayesian Optimization: Searching for the global minima

Erik Karlsson Öhman, Hampus Hansen

August 5, 2025

1 Introduction

This report will analyze the potential energy surface (PES) of an Au surface, and find the position of an additional Au atom (ad-atom) that minimizes the energy using Bayesian optimization with Gaussian processes. The energy of the ad-atom is

$$E(x, y) = E_{\text{ad}}(x, y) - E_{\text{surface}}, \tag{1}$$

where E_{ad} is the energy of the system including the ad-atom, and E_{surface} is the energy of the surface alone. The surface is assumed to be rigid, so to minimize the energy one needs to find the optimal (x, y) position of the ad-atom. Furthermore, the z-dependence has been relaxed [1]. The PES of the (x, y) grid is shown in Figure 1, evaluated using an embedded medium theory calculator (EMT) from the `asap3`-library. Since the system is has many local minima, and the high computational cost associated with evaluating the potential energy, it is appropriate to use Bayesian optimization with Gaussian processes for this problem.

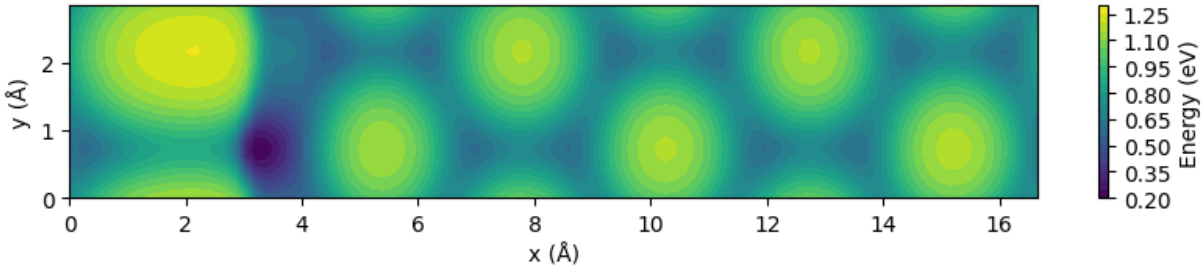


Figure 1: Potential energy surface of the ad-atom of the primitive cell.

The Gaussian processes (GP) is constructed using a RBF kernel and a bias kernel from the `GPy`-library, whose length scale and variance will be discussed later. The kernels, combined with the data from 5 initial samples, constructs the GP model using the `GPRegression()`-function. The coordinates are randomly drawn from uniform distributions, and the energies are calculated using the EMT calculator. The new points are chosen such that they maximize the lower confidence bound (LCB) acquisition function,

$$A(x, y) = -\mu(x, y) + \beta\sigma(x, y), \qquad (x_{\text{new}}, y_{\text{new}}) = \arg\max_{x,y} A(x, y), \tag{2}$$

where β is a hyperparameter. A smaller β results in more exploitation, searching regions with lower energy. A larger β increases the exploration, searching regions with higher uncertainty and avoids getting stuck in local minima. For this new point $(x_{\text{new}}, y_{\text{new}})$, the energy is evaluated using EMT. This new sample is added to the others, and the GP model is retrained. This process is repeated until convergence is reached.

2 Local search

The gradient decent method is implemented using the `scipy.optimize.minimize`-function, with the BFGS algorithm. The local search was initialized by 250 randomly drawn positions. The resulting minima are shown in Figure 2, where $8.4\,\%$ of the samples ended up in the global minimum, whilst the rest got stuck in local minima. This is consistent with the size of the region that is in the vicinity of the global minima, approximately $10\,\%$ of the grid.

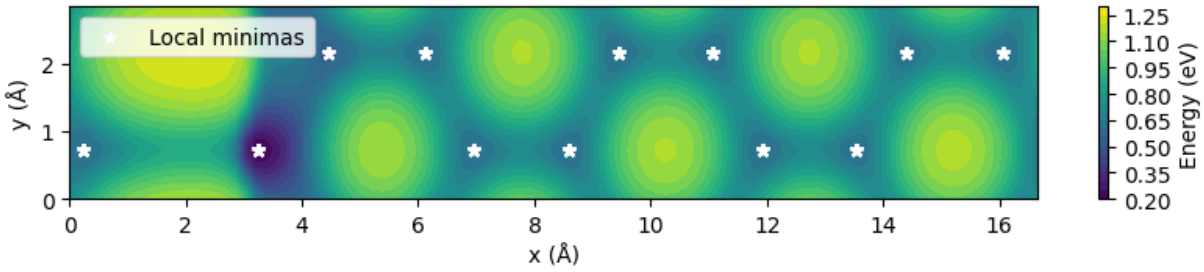


Figure 2: Potential energy surface of the ad-atom with minima found using gradient decent.

3 Bayesian optimization

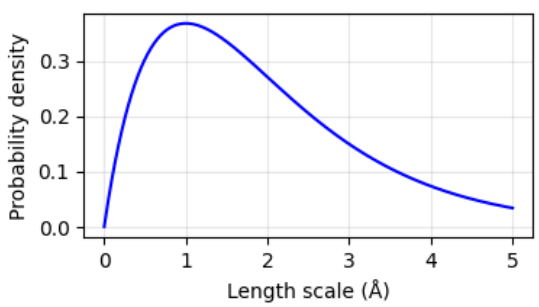
The prior for the characteristic length scale of the system is seen in Figure 3a. It is a gamma distribution

$$p(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{(\alpha-1)} e^{-\lambda x}, \tag{3}$$

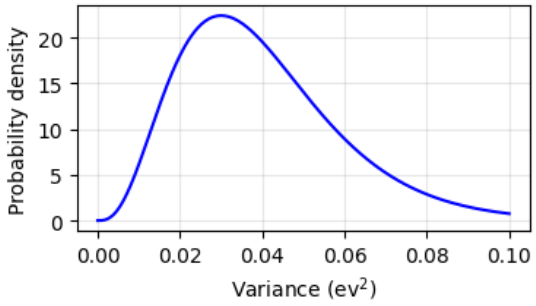
where $\alpha = 2$ is the shape parameter and $\lambda = 1$ is the rate parameter, resulting in a mean length scale of 2, comparable to the scale of the typical distance between two neighboring maxima and minima in the PES from Figure 1. From Figure 2 we see that, for most of the parameter space the value of the energy is in the range of $0.50 - 1.25$ eV. A rough initial estimate would be that this range of energies corresponds to 4 standard deviations. Thus the variance may be estimated to

$$\sigma = \frac{1.25 - 0.25}{4} \implies \sigma^2 = \left(\frac{0.75}{4}\right)^2 \approx 0.04. \tag{4}$$

A gamma distribution for the variance prior distribution with $\alpha = 4$ and $\lambda = 100$, has a mean variance of $0.04\,\text{eV}^2$. This is also comparable to the variance calculated from the explicit evaluation of PES in task 1.



(a) Prior distribution of the length scale.



(b) Prior distribution of the variance.

Figure 3: Prior probability distribution functions of length scale (a) and variance (b).

An analysis of the Bayesian optimization for different choices of the hyperparameter β is presented in Figure 4. Multiple such Bayesian optimizations were performed, and one such representable case is presented below. The smaller βs reach convergence quicker, but they also risk getting stuck in local minima, which is the case for $\beta = 1$. For $\beta \geq 2$ the model almost always converges at the global minima, however, it in general requires more iterations. A conservative parameter choice would thus be $\beta = 2.5$, which increases the exploration slightly, which mitigates the risk of getting stuck in local minima without requiring excessive sampling.

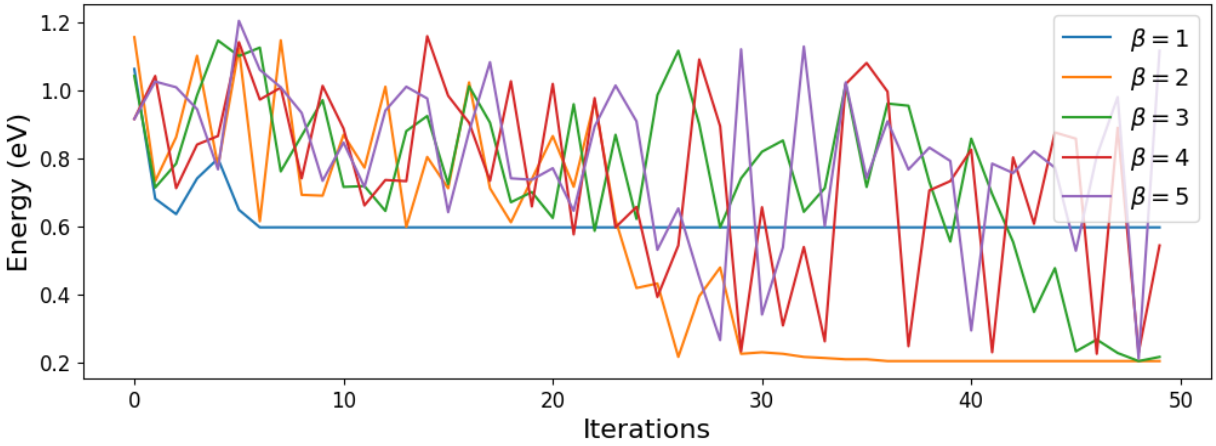
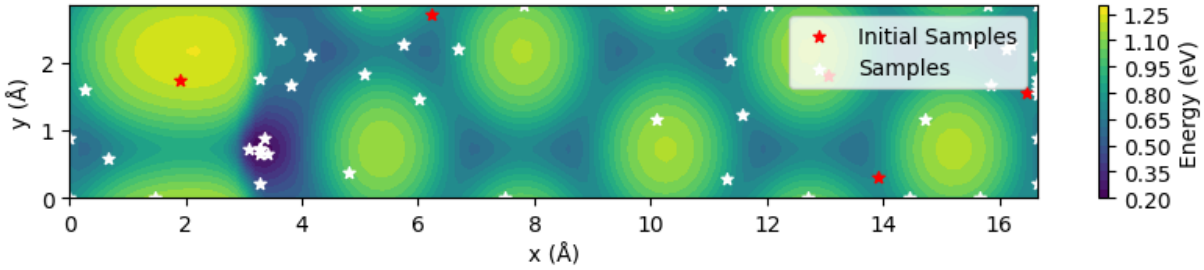


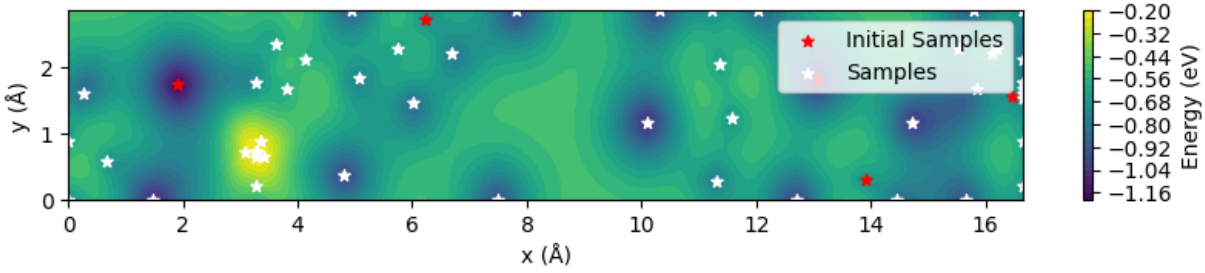
Figure 4: Convergence of the predicted energy for 5 different β.

The sampling sequence of the GP model with $\beta = 2.5$ is seen in Figure 5. The model reached convergence at the global minima after 53 iterations. Figure 5a shows the PES and the GP samples, which are decently spread out and heavily concentrated at the global minimum. Figure 5b shows the samples against the acquisition function at the end of the sampling sequence. Here we see that the global maximum coincides with the maximum of the acquisition function, effectively driving the samples to the minimum. Figure 5c shows the uncertainty σ of the sampling region. The region with the highest uncertainty is the large yellow region, centered around $(x \approx 9\,\text{Å}, y \approx 1\,\text{Å})$. The GP model has not sampled the parameter space here, resulting in high uncertainty. A larger β would cause more exploration, but the model stopped exploration of the parameter space once it converged to the global minimum. The exploration of the parameter space could be considered sufficient since the model found the global minimum, making the region of high uncertainty a non-issue. However, the model could be further improved by dynamically changing β throughout exploration, which could result in more efficient

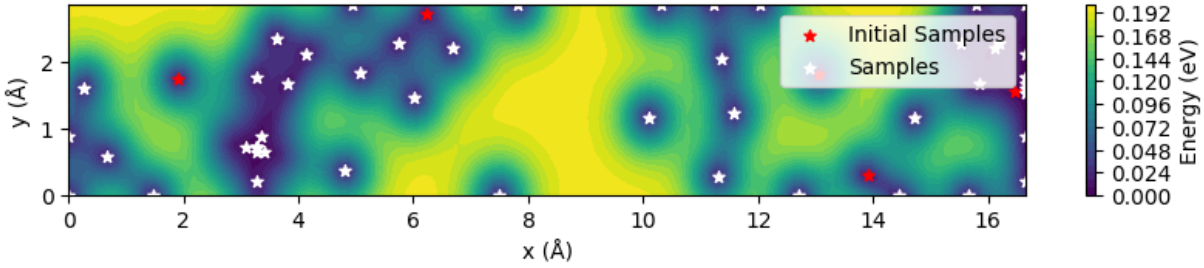
exploration and quicker convergence.



(a) Potential energy surface of the ad-atom with samples from the GP model.



(b) LCB acquisition function on the (x, y) grid.



(c) Uncertainty σ on the (x, y) grid.

Figure 5: Visualization of the potential energy surface (a), acquisition function (b), and uncertainty (c) of the GP model during Bayesian Optimization. Each marker is a sample of the GP model.

4 Transition paths barriers

When creating a general purpose GP model we let $\beta \to \infty$ in our LCB acquisition function, i.e. we effectively get

$$A(x, y) = \sigma(x, y) \tag{5}$$

where we now fully prioritize exploration over exploitation. To initialize the model we use 5 random samples, $(x, y)_{\text{global min}}$ and $(x, y)_{\text{start}} = (11\,\text{Å}, 2.1\,\text{Å})$. To explore how many more samples are needed to create a sufficiently good model we plot the root mean square error (RMSE) between the PES calculated with the EMT model and the PES estimated from the general purpose GP model, as a function of samples. In Figure 6 we can see that 100 additional samples gives an RMSE of $0.093\,\text{eV}$. This is deemed a fair tradeoff between number of samples and model accuracy and is thus used below. Performing another 100 samples only lowers the RMSE by $0.04\,\text{eV}$.

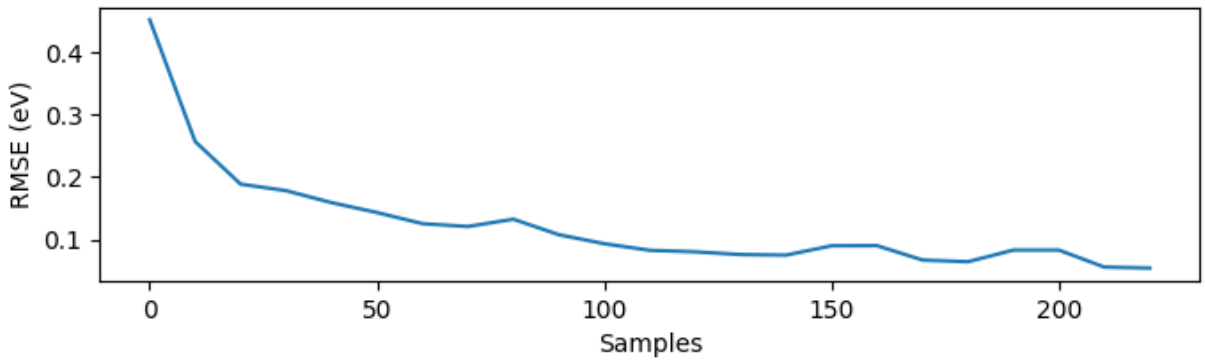


Figure 6: Root mean square error as a function of the number of samples, excluding the 5 starting samples and the start- and endpoint samples.

Estimating the energy along the linear trajectory

$$(x, y) = (x, y)_{\text{start}} + \lambda((x, y)_{\text{global min}} - (x, y)_{\text{start}} \tag{6}$$

where λ varies linearly from 0 to 1, we see in Figure 7 that the EMT model is within one standard deviation of the general purpose GP model that was trained on 100 additional samples. We may note that the uncertainty of the general purpose model does not substantially vary along the path, this is an indicator that we have sampled the parameter space evenly and we have chosen a sufficient number of samples. The model trained in the previous task, with $\beta = 2.5$ does not predict the energy accurately. This is expected since we can see in Figure 5a that there are very few samples in the proximity of the linear path from $(x, y)_{\text{start}}$ to $(x, y)_{\text{global min}}$, thusly the model lacks sufficient data to predict the energy along the path. However at the very end of the trajectory we see that the $\beta = 2.5$-model converges to the value of the EMT-model, which is not the case for the general purpose GP model.

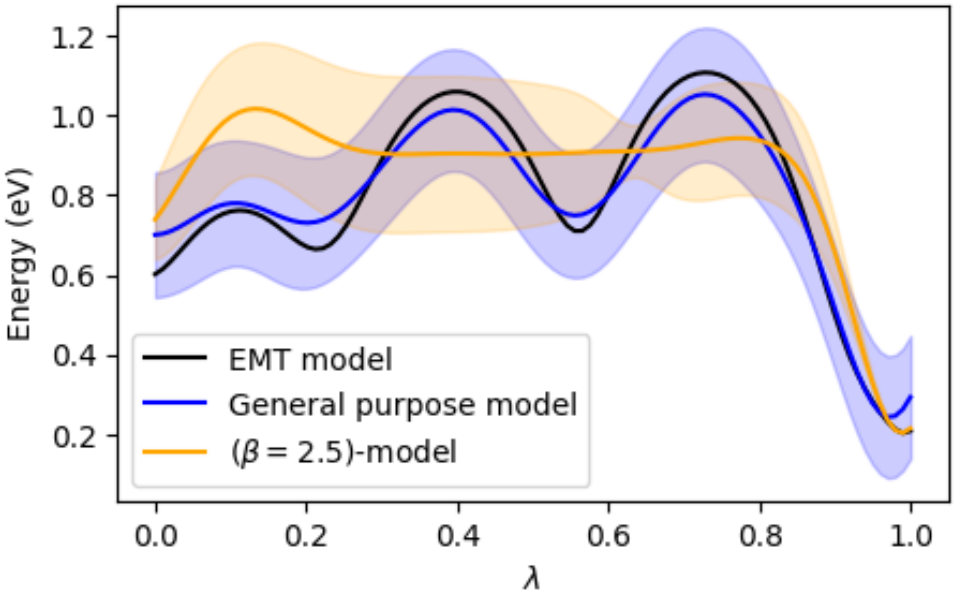


Figure 7: Energy along the linear path from $(x, y)_{\text{start}}$ to $(x, y)_{\text{global min}}$. The black line represents the EMT-calculated energy. The blue line is the general purpose GP model with its $68\,\%$ confidence interval. The orange line is the $\beta = 2.5$-model with its $68\,\%$ confidence interval.

References

[1] P. Erhart, *Advanced Simulation and Machine Learning: Project 2b*. Gothenburg, Sweden: Chalmers University of Technology, Nov. 20, 2024.