Machine-learning-based selective sampling procedure for identifying the low-energy region in a potential energy surface: A case study on proton conduction in oxides

Kazuaki Toyoura, Daisuke Hirano, Atsuto Seko, Motoki Shiga, Akihide Kuwabara, Masayuki Karasuyama, Kazuki Shitara, and Ichiro Takeuchi

A machine learning-based selective sampling procedure for identifying the low energy region in a potential energy surface: a case study on proton conduction in oxides

Kazuaki Toyoura,[1, *] Daisuke Hirano,[2, †] Atsuto Seko,[3, 4, 5] Motoki Shiga,[6] Akihide Kuwabara,[7] Masayuki Karasuyama,[2, 5] Kazuki Shitara,[4, 7] and Ichiro Takeuchi[2, 5, ‡]

[1]*Department of Materials Science and Engineering, Nagoya University, Nagoya 464-8603, Japan*
[2]*Department of Engineering, Nagoya Institute of Technology, Nagoya 466-8555, Japan*
[3]*Department of Materials Science and Engineering, Kyoto University, Kyoto 606-8501, Japan*
[4]*Center for Elements Strategy Initiative for Structure Materials (ESISM), Kyoto University, Kyoto 606-8501, Japan*
[5]*Center for Materials Research by Information Integration, National Institute for Materials Science, Tsukuba 305-0047, Japan*
[6]*Department of Electrical, Electronic and Computer Engineering, Gifu University, Gifu, 501-1193, Japan*
[7]*Nanostructures Research Laboratory, Japan Fine Ceramics Center, Nagoya, 456-8587, Japan*
(Dated: February 4, 2016)

In this paper, we propose a selective sampling procedure to preferentially evaluate a potential energy surface (PES) in a part of the configuration space governing a physical property of interest. The proposed sampling procedure is based on a machine learning method called the *Gaussian process (GP)*, which is used to construct a statistical model of the PES for identifying the region of interest in the configuration space. We demonstrate the efficacy of the proposed procedure for atomic diffusion and ionic conduction, specifically the proton conduction in a well-studied proton-conducting oxide, barium zirconate ($BaZrO_3$). The results of the demonstration study indicate that our procedure can efficiently identify the low-energy region characterizing the proton conduction in the host crystal lattice, and that the descriptors used for the statistical PES model have a great influence on the performance.

I. INTRODUCTION

The concept of a potential energy surface (PES) is of great importance for a fundamental understanding of physical phenomena and properties, such as molecular and lattice vibrations, thermodynamic properties, chemical reactions, atomic diffusion and ionic conduction [1–6]. A PES is generally described as a function of $3n$ degrees of freedom in an n-atom system, but it is often simplified by focusing only on reduced degrees of freedom that govern the physical property of interest [7–9]. In this simplification, the rest of the degrees of freedom are effectively incorporated into the reduced configuration space as the minimized energy or statistical average. For example, when tracing the movement of an diffusing atom in a solid, the host lattice structure is optimized with respect to the energy, so that the movements of the other atoms forming the host crystal lattice are implicitly treated [8, 9]. This approach is useful for evaluating physical properties mainly governed by the motions of a few atoms, such as atomic diffusion, ionic conduction, and elementary steps of chemical reactions.

In the simplified approach, the entire PES is evaluated for the reduced degrees of freedom, generally by density functional theory (DFT) calculations. A fine grid is usually introduced in the reduced configuration space to evaluate the potential energies (PEs) at uniformly-distributed points in the configuration space. The most accurate method is for the PEs at all the grid points to be fully computed with the host lattice structure optimization, which requires that a great amount of computational resources are devoted. However, one should note that the entire PES is not necessarily required for understanding physical phenomena and properties. In the case of atomic diffusion and ionic conduction, a diffusion or a conduction process is fully characterized by a connected low-energy migration pathway throughout the host lattice, meaning that we need not to explore grid points with high PEs irrelevant to the migration pathway [10, 11]. Therefore, it is sufficient to know only a region exhibiting low energy, called the *low PE region*, for finding the migration pathway.

In this paper, we propose a selective sampling procedure on the basis of a machine learning technique for evaluating all points in the low PE region containing the migration pathway without evaluating the entire PES. Here we introduce a statistical PES model to sample points in the low PE region efficiently and to ensure that all points in the low PE region have already been sampled without evaluating the entire PES. In particular, the latter is difficult to achieve without using a statistical model. This procedure is generally applicable to any case of finding points in a region of interest, if the region can be mathematically defined. For example, local minima and maxima, ridge and valley lines, and saddle points in a PES are defined mathematically using the PE gradients, which can also be specified as a region of interest to be sampled

selectively. In addition, beyond such PES evaluation, it can also be used for exploring materials with an over- or sub-threshold physical property if an appropriate set of descriptors is found for the physical property of interest. For demonstrating the performance of our selective sampling procedure, we apply it to proton conduction in a well-studied proton-conducting oxide, barium zirconate ($BaZrO_3$) [12–15]. Specifically, a fine grid is introduced in the crystal lattice of the host oxide, and several sampling procedures are compared in terms of efficiency for finding all the grid points with PEs less than a given PE threshold.

Note that the proposed procedure is based on a different concept than those of previously reported machine-learning-based material science studies [16–33]. Specifically, the goal of our procedure is NOT estimating the entire PES from a sparse sampling using regression methods [17, 19, 26]. In the present study, machine-learning techniques are used to identify all points related to the partial PES governing a physical property of interest. Consequently, the present procedure identifies the complete partial PES. Although a hybrid method of the proposed and previous procedures should be more efficient than the present approach for evaluating a reasonable partial PES, this paper focuses on the procedure for finding all points related to the partial PES for simplicity.

II. GAUSSIAN PROCESS MODEL-BASED SELECTIVE SAMPLING PROCEDURE

In this section, we present the proposed selective sampling procedure in detail. Although the procedure is fairly general and can be used in a variety of problems, we describe it here in the context of low PE region detection problems. Given a fine set of grid points in the configuration space, the goal is to efficiently identify the subset of the grid points at which the PEs are relatively low. The procedure has the following three key features. (1) First, a statistical model of the PES is developed as a *Gaussian process (GP)* [34, 35]. The model is iteratively updated by repeating (i) selectively sampling a point at which the PE is predicted to be low, and (ii) updating the model based on the newly calculated PE value at the sampled point. (2) Second, the statistical PES model is used for identifying the subset of grid points at which the PEs are relatively low. We introduce a novel selection criterion for this purpose, because the single global minimum or maximum point (not a subset) has been targeted in GP applications so far. (3) Finally, the procedure allows us to estimate how many low PE points remain unsampled, i.e., lets us know when the sampling should be stopped.

These features are made possible by exploiting an advantage of GP that it provides not only the predicted PE value but also the uncertainty at each grid point. Using this property of GP modeling, we develop a new sampling procedure which enables us to sample grid points preferentially from the low PE region, and tells us when we

should stop sampling. FIG.1 illustrates the sequences of the selective sampling by taking a one-dimensional synthetic PES as an example, in which 9 grid points in the low PE region (within the blue bar) should be selectively sampled from all the points (50 points). Roughly speaking, at each step, a point most likely to be located in the low PE region is sampled in the light of the predicted PEs (red solid curve) and the uncertainties (pale red area). In early steps, the predicted PEs are quite uncertain and have large discrepancies from the true PES (black solid curve), resulting in selecting grid points whose uncertainties are large. As the sampling proceeds, the predicted PE curve gradually approaches the true one with decreasing uncertainty, so that grid points in the low PE region are selectively sampled in later steps. The uncertainty in the GP model is useful also in the judgment to stop the sampling. The stopping criterion should be determined on the basis of the existence probability of unsampled low PE points, for which the information on the uncertainty is indispensable in addition to the predicted PEs.

In the remaining part of this section, we first present the problem setup in §II A, and then describe the details of the three key features in §II B, §II C, and §II D.

A. Problem setup

Suppose that there are N grid points, $i = 1, \ldots, N$, in the configuration space. For each point, we denote by E_i the potential energy (PE). In order to investigate a target physical property such as proton conductivity, we need to identify a region whose PEs are *relatively* low. Using a parameter $0 < \alpha < 1$, we define such a *relatively* low PE region as the set of αN points at which the PEs are lower than the PEs at the other $(1-\alpha)N$ points. We refer to such a subset of points as the "low PE region", and the remaining set of points as the "high PE region". The goal of the problem is to identify all the points in the low PE region as efficiently as possible. For simplicity, we assume that α is pre-specified, but it can be adaptively determined as we demonstrate in § V B.

The problem of identifying the low PE region is formulated as follows. Let θ_α represent the threshold of the low PE region, and define

$$\mathcal{P}_\alpha := \{i \in \{1, \ldots, N\} \mid E_i < \theta_\alpha\},$$
$$\mathcal{N}_\alpha := \{i \in \{1, \ldots, N\} \mid E_i \geq \theta_\alpha\}.$$

Then, the task is formally stated as the problem of identifying all the points in $\mathcal{P}_\alpha$. Borrowing the terminology in statistics, we refer to the points in $\mathcal{P}_\alpha$ and $\mathcal{N}_\alpha$ as *positive* and *negative* points, respectively.

Note that we do not know $\mathcal{P}_\alpha$, $\mathcal{N}_\alpha$ and θ_α unless we actually compute the PEs at all the grid points. Therefore, during the sampling process, these quantities are estimated based on the PEs at the points sampled in earlier steps. We denote our estimates of positive and negative sets as $\hat{\mathcal{P}}_\alpha$ and $\hat{\mathcal{N}}_\alpha$, respectively, i.e., the former
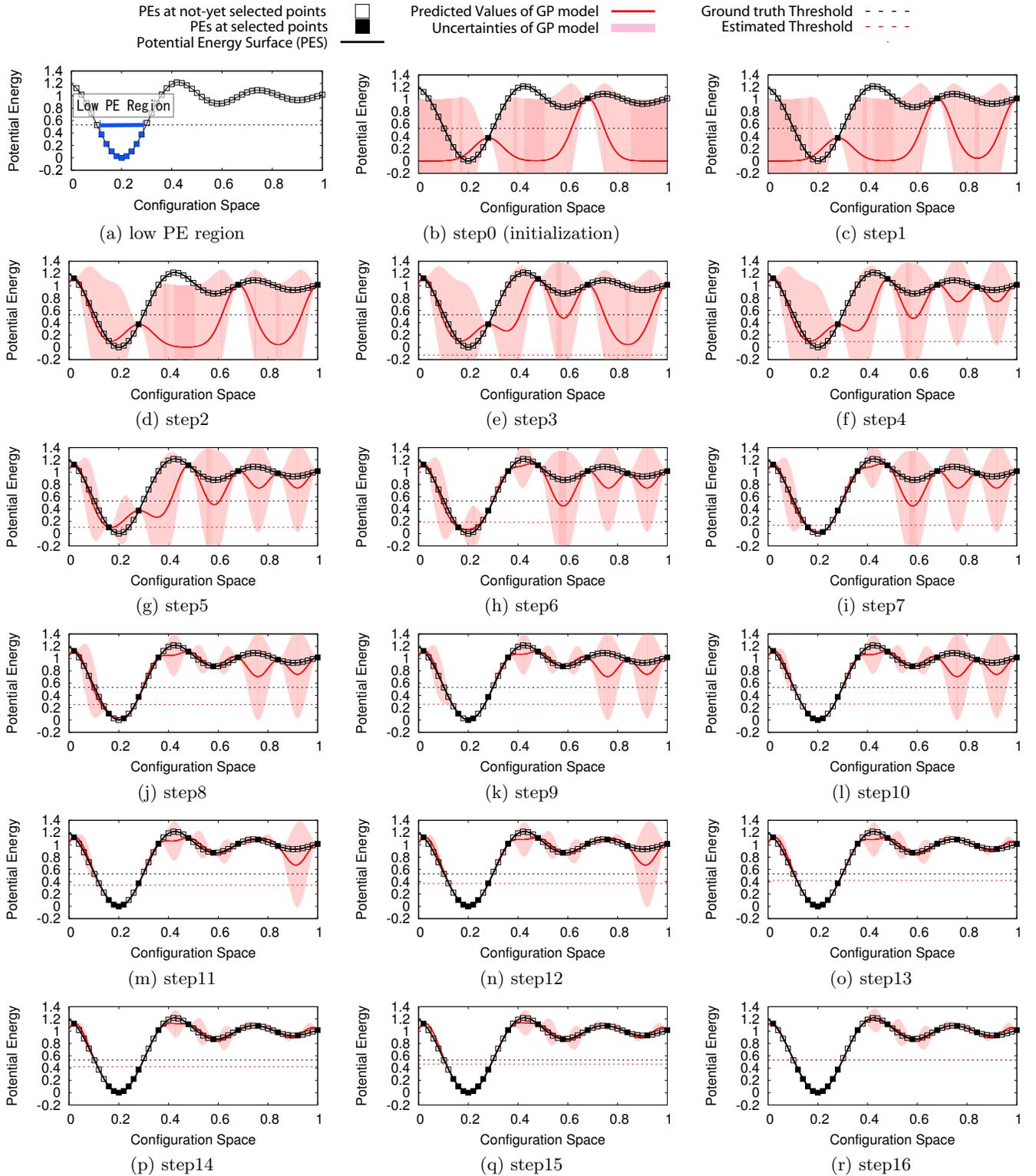
FIG. 1. Schematic illustration of the proposed selective sampling procedure in one-dimensional configuration space with synthetic data. In each plot, the X and Y-axes represent the configuration space and the PEs, respectively. The PES and its $N = 50$ points are shown as a black solid curve and squares, respectively. In plot (a), the low PE region with $\alpha = 0.20$ is represented by a blue bar. The goal in this example is to efficiently identify and evaluate the PEs at the 9 points within this blue bar region. Plot (b) indicates the initialization step, where two points (filled squares) are randomly selected and their PEs are evaluated. The remaining 16 plots (plots (c) to plot (r)) indicate steps 1 to 16 of the procedure. At each step, a point is selected and the PE is evaluated by using the GP-based statistical model, where the points selected so far are shown as filled squares. In each plot, the predicted PEs and their uncertainties given by the GP-based statistical model are depicted by a red solid curve and shaded region, respectively. In addition, the true and the estimated threshold values θ_α and $\hat{\theta}_\alpha$ are shown as black and red broken lines, respectively. In this example, the 9 points in the low PE region are successfully identified and the threshold is also correctly estimated at step 16 (plot (r)).

indicates the set of points at which the PEs have already been sampled and evaluated in earlier steps, while the latter indicates the set of points at which the PEs have not yet been evaluated. The proposed selective sampling procedure can be interpreted as the process of sequentially updating these two sets of points. Specifically, we start from $\hat{\mathcal{P}}_\alpha = \emptyset$ and $\hat{\mathcal{N}}_\alpha = \{1, \ldots, N\}$, and then update the two sets as $\hat{\mathcal{P}}_\alpha \leftarrow \hat{\mathcal{P}}_\alpha \cup \{i'\}$ and $\hat{\mathcal{N}}_\alpha \leftarrow \hat{\mathcal{N}}_\alpha \setminus \{i'\}$ where i' is the sampled point in the step. When the stopping criterion is satisfied, $\hat{\mathcal{P}}_\alpha$ is expected to contain all the points in $\mathcal{P}_\alpha$ with high probability. We also denote the estimate of θ_α as $\hat{\theta}_\alpha$ (see §II C for how to estimate θ_α).

B. Gaussian process (GP) models

We adopt GP [34, 35] as a choice of the statistical model of the PES. Using a GP model, the potential energy E_i is represented in the following form:

$$E_i \sim N(\mu_i, \sigma_i^2), \ i = 1, \ldots, N, \tag{1}$$

where $N(\mu_i, \sigma_i^2)$ denotes the Normal distribution with mean μ_i and variance σ_i^2.

A GP model is a sort of regression model. Let us consider a d-dimensional vector of descriptors for each point and denote the vector as $\boldsymbol{\chi}_i \in \mathbb{R}^d$ for $i = 1, \ldots, N$. Then, the mean and variance of the PE at the ith point are represented as functions of $\boldsymbol{\chi}_i$ as in (4) and (5) below. In a GP model, a so-called *kernel function* $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is employed. For two different points indexed by i and j, $k(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)$ is roughly interpreted as a similarity between these two points. One of the most commonly used kernel functions is the *RBF kernel*

$$k(\boldsymbol{\chi}, \boldsymbol{\chi}') = \gamma_1^2 \exp(-\|\boldsymbol{\chi} - \boldsymbol{\chi}'\|^2 / 2\gamma_2^2), \tag{2}$$

where $\gamma_1, \gamma_2 > 0$ are tuning parameters, and $\|\cdot\|$ represents the L_2 norm. Furthermore, for n points indexed by $1, \ldots, n$, let $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ be the so-called *kernel matrix* defined as

$$\boldsymbol{K} := \begin{bmatrix} k(\boldsymbol{\chi}_1, \boldsymbol{\chi}_1) & \cdots & k(\boldsymbol{\chi}_1, \boldsymbol{\chi}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{\chi}_n, \boldsymbol{\chi}_1) & \cdots & k(\boldsymbol{\chi}_n, \boldsymbol{\chi}_n) \end{bmatrix}. \tag{3}$$

Then, for any points in the configuration space whose descriptor vector is represented as $\boldsymbol{\chi} \in \mathbb{R}^d$, the GP model provides the predictive distribution of its PE in the form of a Normal distribution $N(\mu(\boldsymbol{\chi}), \sigma^2(\boldsymbol{\chi}))$. Here, the mean function $\mu : \mathbb{R}^d \to \mathbb{R}$ is given as

$$\mu(\boldsymbol{\chi}) := \boldsymbol{\kappa}(\boldsymbol{\chi})^\top \boldsymbol{K}^{-1} \boldsymbol{E}, \tag{4}$$

where $\boldsymbol{\kappa}(\boldsymbol{\chi}) := [k(\boldsymbol{\chi}, \boldsymbol{\chi}_1), \ldots, k(\boldsymbol{\chi}, \boldsymbol{\chi}_n)]^\top$ and $\boldsymbol{E} := [E_1, \ldots, E_n]^\top$, while the variance function $\sigma^2 : \mathbb{R}^d \to \mathbb{R}$ is given as

$$\sigma^2(\boldsymbol{\chi}) := k(\boldsymbol{\chi}, \boldsymbol{\chi}) - \boldsymbol{\kappa}(\boldsymbol{\chi})^\top \boldsymbol{K}^{-1} \boldsymbol{\kappa}(\boldsymbol{\chi}). \tag{5}$$

At each step of the procedure, we fit a GP model of PES based on $\{(\boldsymbol{\chi}_i, E_i)\}_{i \in \hat{\mathcal{P}}_\alpha}$, which is the set of points that have been selected and whose PEs have been already evaluated by DFT calculations in earlier steps.

C. Selection criterion for low PE region identification

Given a GP model in the form of (1) for each point, the subsequent task is to select the point at which the PE is most likely to be lower than the estimated threshold $\hat{\theta}_\alpha$ (we will discuss how to estimate the threshold later). In this task, we can borrow some techniques developed in the context of *Bayesian optimization* [36, 37], which has been used for minimization or maximization of an unknown function. In the Bayesian optimization literature, there are two main options that can be adapted to our task. The first option is to select the point at which the probability that the PE is lower than $\hat{\theta}_\alpha$ is maximized, which is called *probability of improvement*, and formulated as

$$i' := \arg\max_{i \in \hat{\mathcal{N}}_\alpha} \ \Phi(\hat{\theta}_\alpha; \mu(\boldsymbol{\chi}_i), \sigma^2(\boldsymbol{\chi}_i)), \tag{6}$$

where $\Phi(\cdot; \mu, \sigma^2)$ is the cumulative distribution function of $N(\mu, \sigma^2)$. The second option is *expected improvement*, which is similarly formulated as

$$i' := \arg\min_{i \in \hat{\mathcal{N}}_\alpha} \ \int_{-\infty}^{\hat{\theta}_\alpha} E \ \phi(E; \mu(\boldsymbol{\chi}_i), \sigma^2(\boldsymbol{\chi}_i)) dE, \tag{7}$$

where $\phi(\cdot; \mu, \sigma^2)$ is the probability density function of $N(\mu, \sigma^2)$. In the simulation studies conducted in §IV, we used (7). In our experience, there is little difference in the performances between the choices of (6) and (7).

In order to obtain an estimate $\hat{\theta}_\alpha$ of the threshold θ_α, let us consider the following contingency table

	$\mathcal{P}_\alpha$	$\mathcal{N}_\alpha$
$\hat{\mathcal{P}}_\alpha$	#TP$(\hat{\theta}_\alpha)$	#FP$(\hat{\theta}_\alpha)$
$\hat{\mathcal{N}}_\alpha$	#FN$(\hat{\theta}_\alpha)$	#TN$(\hat{\theta}_\alpha)$

$$\tag{8}$$

where TP, FP, FN, and TN stand for true positive, false positive, false negative, and true negative, respectively, and the notation # indicates the number of the event. These numbers for four events can be rephrased as

- #TP: The number of sampled points in the low PE region.

- #FP: The number of sampled points in the high PE region.

- #FN: The number of not-yet sampled points in the low PE region.

- #TN: The number of not-yet sampled points in the high PE region.

Note that, in (8), these numbers depend on $\hat{\theta}_\alpha$. Remembering that $\mathcal{P}_\alpha/(\mathcal{P}_\alpha+\mathcal{N}_\alpha) = \alpha$, the following relationship should be maintained

$$\frac{\#\mathrm{TP}(\hat{\theta}_\alpha) + \#\mathrm{FN}(\hat{\theta}_\alpha)}{N} = \alpha. \tag{9}$$

Since we have already evaluated E_i for $i \in \hat{\mathcal{P}}_\alpha$, we can simply obtain

$$\#\mathrm{TP}(\hat{\theta}_\alpha) = \sum_{i \in \hat{\mathcal{P}}_\alpha} I(E_i < \hat{\theta}_\alpha), \tag{10}$$

where $I(\cdot)$ is the indicator function, defined by $I(z) = 1$ if z is true and $I(z) = 0$ if z is false. On the other hand, we need to estimate $\#\mathrm{FN}(\hat{\theta}_\alpha)$ based on the statistical model (1) because we do not know E_i for $i \in \hat{\mathcal{N}}_\alpha$:

$$\#\mathrm{FN}(\hat{\theta}_\alpha) \simeq \#\hat{\mathrm{FN}}(\hat{\theta}_\alpha) := \sum_{i \in \mathcal{N}_\alpha} \Phi(\hat{\theta}_\alpha; \mu(\boldsymbol{\chi}_i), \sigma^2(\boldsymbol{\chi}_i)). \tag{11}$$

The estimate of the threshold $\hat{\theta}_\alpha$ is determined in each step so that it satisfies the requirement (9) where the quantities in the left-hand side are given by (10) and (11).

D. How to assess the sampled points

When the sampling is stopped, $\hat{\mathcal{P}}_\alpha$ should ideally contain all the points in $\mathcal{P}_\alpha$, i.e., $\hat{\mathcal{P}}_\alpha \supseteq \mathcal{P}_\alpha$. As we can easily notice from the contingency table in (8), this requirement can be rewritten as $\#\mathrm{FN}(\hat{\theta}_\alpha) = 0$. This indicates that the estimated false negative rate (FNR) defined as

$$\hat{\mathrm{FNR}} := \frac{\#\hat{\mathrm{FN}}(\hat{\theta}_\alpha)}{\#\mathrm{TP}(\hat{\theta}_\alpha) + \#\hat{\mathrm{FN}}(\hat{\theta}_\alpha)} \tag{12}$$

can be used for assessing the *badness* of the sampled points. $\hat{\mathrm{FNR}}$ in (12) can be interpreted as the proportion of points for which the PEs have yet to be evaluated. At each step, we compute $\#\mathrm{TP}(\hat{\theta}_\alpha)$ by (10) and estimate $\#\hat{\mathrm{FN}}(\hat{\theta}_\alpha)$ by (11). Then, the sampling is terminated if $\hat{\mathrm{FNR}}$ is close to zero (e.g., 10^{-6}).

III. PROTON PES IN BARIUM ZIRCONATE

A. Grid points

To show the performance of the selective sampling procedure in the present study, we evaluated the PEs for all the grid points, i.e., the entire PES using DFT calculations with structure optimization. Firstly, grid points are introduced in the host crystal lattice of $BaZrO_3$. FIG.2 shows the crystal structure of $BaZrO_3$ (space group: $Pm\bar{3}m$ (221)) and its asymmetric unit satisfying $0 \leq x, y, z \leq 0.5$, $y \leq x$ and $z \leq y$ where x, y and z

denote the three-dimensional fractional coordinates of a proton introduced into the host lattice. The lattice constant a was experimentally reported to be 4.19-4.20 Å [38, 39], which is reasonably consistent with our calculated value (4.23 Å) using the computational condition as will be described later. Ba, Zr and O ions occupy $1a$, $1b$, and $3c$ sites, respectively, in the case of the origin setting shown in FIG.2.
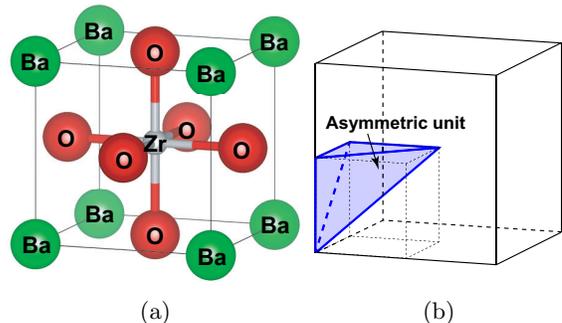
FIG. 2. (a) Crystal structure of $BaZrO_3$. (b) Asymmetric unit in the unit cell of $BaZrO_3$.

Considering a $40 \times 40 \times 40$ grid in the unit cell (grid interval is nearly equal to 0.1 Å), there are 64,000 grid points in total. Reflecting the high crystallographic symmetry of $BaZrO_3$, the number of grid points in the asymmetric unit shown in the right plot of FIG.2 is reduced to 1771 points. In other words, these PEs enable us to construct the entire three-dimensional PES using symmetry operations. Among them, three points exactly coincide with a Ba ion, a Zr ion, and an O ion. Further removing those three points, the remaining grid points are reduced to $N = 1768$ points.

B. Procedure of DFT calculation

DFT calculations for the PES evaluation in $BaZrO_3$ were based on the projector augmented wave (PAW) method as implemented in the VASP code [40–43] The generalized gradient approximation (GGA) parameterized by Perdew, Burke, and Ernzerhof was used for the exchange-correlation term [44]. The $5s$, $5p$ and $6s$ orbitals for barium, $4s$, $4p$, $5s$ and $4d$ for zirconium, $2s$ and $2p$ for oxygen, and $1s$ for hydrogen were treated as valence states. The plane-wave cutoff energy was set to be 400 eV. A supercell consisting of $3 \times 3 \times 3$ unit cells (135 atoms) was used, with a $2 \times 2 \times 2$ mesh for the k-point sampling, in which the atomic positions only in the limited region corresponding to $2 \times 2 \times 2$ unit cells around the introduced proton were optimized with the other atoms and the proton fixed. The atomic positions were optimized until the residual forces converged to less than 0.02 eV/Å.

C.   Evaluated PES

FIG.3 shows the evaluated PES of a proton in a low PE region for BaZrO$_3$. All points in the low PE region with $\alpha = 0.2$ show a PE of less than 0.30 eV. The blue spheres bonding to single O ions in the figure denote the most stable proton sites, which are located ~ 1 Å from the O ions. The OH distance is equivalent to that in water, indicating that protons in BaZrO$_3$ are stabilized by forming an OH bond. There are four equivalent proton sites per O ion, which are connected by low PE points around the O ion. This is the rotational path around the O ion consisting of four equivalent quarter rotational paths, whose potential barrier is 0.18 eV. The strong OH bonding is always maintained during the proton rotation, leading to the low potential barrier. On the other hand, the hopping path connecting adjacent rotational orbits is located at the periphery of the edges of the ZrO$_6$ octahedra. The bonding with adjacent O ions is the key also in the proton hopping, where not the OH bond but the O-H-O bond (like the hydrogen bond) is formed at the transition state. The two O ions involving the proton transfer gradually approach each other to give and take the migrating proton, and the O-O distance is finally reduced to 2.4 Å at the transition state from 2.9 Å at the initial state. The proton is just located on the middle plane between the two O ions at the transition state, and the two OH bond lengths are as short as 1.24 Å. As a result, the calculated potential barrier of the hopping path is not so high (0.25 eV), although it is slightly higher than that of the rotational path. The large host lattice relaxation to reduce the O-O distance is a possible origin of the relatively higher potential barrier. The calculated potential barrier of the hopping path
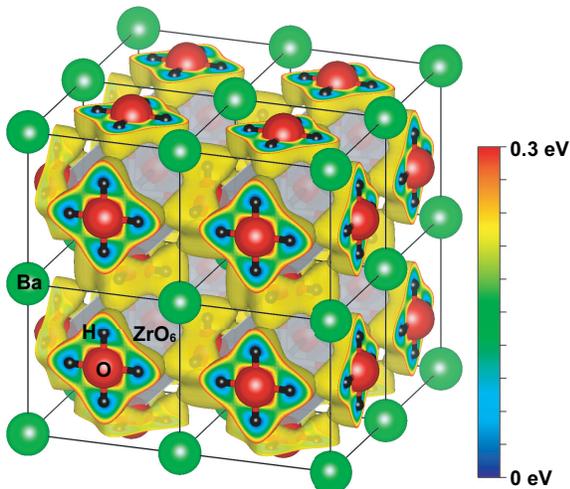
is 0.25 eV, which is higher than that of the rotational path. The two kinds of paths form a three-dimensional proton-conducting network throughout the crystal lattice, meaning that protons migrate over a long range by repeated rotation and hopping. The conduction mechanism of the rotation and hopping has been previously studied based on DFT analysis with the nudged elastic band (NEB) method [12, 45, 46]. The reported potential barriers of the rotation and the hopping are in the range of 0.10-0.18 eV and 0.23-0.27 eV, respectively, which are in good agreement with our results. kThe scattering of the reported potential barriers are probably due to the difference in the computational condition. Particularly, these three reports employed different manners for charge compensation of the computational cell with a positively-charged proton, i.e., a negatively-charged In^{3+} dopant on a Zr^{4+} site [45], a homogeneous background charge [12], and an electron [46]. The homogeneous background charge was here employed, leading to the comparable potential barriers to those in Ref. [12].

IV.   RESULTS OF SELECTIVE SAMPLINGS

In this section, we illustrate the advantage of the proposed selective sampling procedure through a simulation study in which we apply it to a low PE region identification problem for a proton in BaZrO$_3$.

A.   Setup of the simulation study

We compared the performances of several sampling methods for $\alpha = 0.05$, 0.1 and 0.2. Specifically, we compared the following seven methods:

- `GP1(xyz)`

- `GP2(xyz + 1st NNs)`

- `GP3(xyz + decaying NNs)`

- `GP4(xyz + prePES)`

- `random`

- `prePES`

- `ideal`

The first four methods are the proposed GP-based selective sampling procedure with different descriptors. In `GP1`, we used the three-dimensional coordinates x_i, y_i, z_i in the host crystal lattice as the descriptors of the ith point (denoted as "`xyz`"). In `GP2`, we also used the first-nearest-neighbor (NN) distances to Ba, Zr and O atoms from each point as additional descriptors (denoted as "`1st NN`"). In `GP3`, a weighted sum of the distances to the 1st NN atoms, 2nd NN atoms, and so on, are also incorporated as additional descriptors, where the weights are decaying to zero as the atom goes away from the

FIG. 3. PES of a proton in the low PE region ($\alpha = 0.2$) obtained by DFT calculations with structural optimization for BaZrO$_3$. The low PE region ($\alpha = 0.2$) at which the PEs are lower than 0.30 eV is surrounded by yellow surfaces.

TABLE I. Numbers of sampling steps required to identify all the grid points in the low PE region. The averages and the standard deviations over 10 runs with different random seeds are shown. Note that the results of `prePES` are obtained by a single run because there is no random component in `prePES`.

	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$
GP1	1488.5 ± 239.4	1576.1 ± 36.9	1539.6 ± 31.2
GP2	1294.1 ± 180.9	1266.9 ± 167.9	1269.4 ± 100.3
GP3	511.7 ± 30.4	617.6 ± 49.7	875.2 ± 60.1
GP4	114.3 ± 1.3	235.5 ± 1.0	394.1 ± 5.2
random	1749.6 ± 15.0	1757.1 ± 11.3	1762.2 ± 3.9
prePES	592	1222	1479
ideal	89	177	354

point (denoted as "decaying NNs"). In GP4, we used *preliminary PES* (denoted as "prePES") as an additional descriptor. Here, we mean by preliminary PES a *rough but quick* approximation of the PES obtained by using less accurate but more efficient computational methods. Specifically, we used PE values at all the N points obtained by single-point DFT calculations using a smaller supercell consisting of $2 \times 2 \times 2$ unit cells with a $2 \times 2 \times 2$ k-point sampling. In GP4, GP model is used for modeling the difference between the actual PES and the preliminarily PES [33].

The method `random` indicates a naive random sampling, where a point is selected at each step uniformly at random. The method `prePES` indicates a selective sampling method based only on a preliminary PES. Specifically, points are sequentially selected in the ascending order of the preliminary PEs obtained by single-point DFT calculations. Finally, `ideal` indicates the best possible ideal sampling method which is only possible when the actual PEs at all the points are known in advance.

In GP1 to GP4, we must select two points at random for initializing the GP model. Thus, we report the average and the standard deviation over 10 runs with different random seeds. The tuning parameters of the GP models were set as $\gamma_1 = \gamma_2 = 0.5$. According to our preliminary experiments (not shown), the performances were not very sensitive to the choices of those tuning parameters.

B. Performance comparison

FIG.4 and Table I compare the efficiencies of the seven sampling methods. In each of the three plots (corresponding to the $\alpha = 0.05$, 0.1, and 0.2 cases), the number of points successfully sampled from the low PE region ($\#(\mathrm{TP})$) is plotted as a function of the number of PE evaluations by DFT calculations ($\#(\mathrm{TP}) + \#(\mathrm{FP})$).

The results for the four different GP-based methods (GP1 to GP4) indicate the importance of the choice of descriptors. When only the three-dimensional coordi-

nate information (`xyz`) is used, the performances were only slightly better than `random`. For example, when $\alpha = 0.20$, GP1 required 1539.6 DFT calculations in average until all the points in the low PE region were successfully identified. The results on GP2 and GP3 were better than GP1, suggesting that adding appropriate descriptors are generally advantageous. The results on GP4 indicate that a preliminary PES is highly helpful as a descriptor. However, the results on `prePES` suggest that the preliminary PES alone is not sufficient for efficiently identifying the low PE region. In §IV C, we discuss the importance of the preliminary PES in more detail.

FIG.5 demonstrates the differences between the sampling sequences of the GP1, GP4, `prePES`, `random` and `ideal` methods (we omit those of GP2 and GP3 due to space limitations). In GP1, many points in the high PE region (non-low PE region) were mistakenly selected. In GP4, the points in the low PE region were preferentially selected, and only a small number of points were mistakenly selected from the high PE region. In `prePES`, the method preferentially selected the points in the low PE region, but it failed to find all of them. Note that the sampling sequence of GP4 looks almost identical to that in the `ideal` sampling, despite the low PE region being unknown beforehand. This indicates that the GP model in GP4 could successfully estimate the PES in the low PE region.

C. Importance of preliminary PES

The results in FIG.4 and Table I indicate that the preliminary PES obtained by single-point DFT calculations is highly valuable as a descriptor when it is used along with the three-dimensional coordinates (`xyz`) in GP modeling. However, using the preliminary PES alone is not sufficient for identifying the low PE region in the method `prePES`, which was only slightly better than `random`. Although the sampling curves of `prePES` almost overlap with the ideal sampling curves in earlier steps, they gradually deviate as the sampling proceeds. Eventually, the total numbers of sampling steps required for finding all the points in the low PE region were as large as 592, 1222, and 1479 points for the $\alpha = 0.05$, 0.10, and 0.20 cases respectively, which are 6.7-fold, 6.9-fold, and 4.2-fold worse than the `ideal` sampling cases. This inefficiency of `prePES` is ascribed to the relationship between the DFT calculation with structural optimization and the single-point DFT calculation. This can be clearly seen in the rank correlation between the actual and preliminary PEs shown in FIG.6. FIG.6 shows the rank correlation for the $\alpha = 0.20$ case, where points with low PEs are located below the horizontal broken line. Since the `prePES` sampling method selects the points in the increasing order of the preliminary PEs, meaning that the points are selected from left to right in FIG.6, most of the N grid points (all points located in the shaded region) must be sampled for selecting all the points in the low PE region.
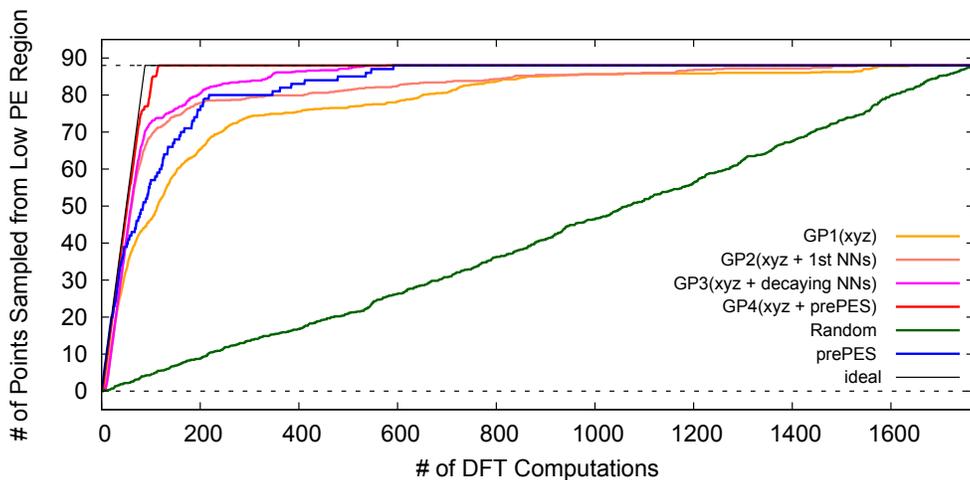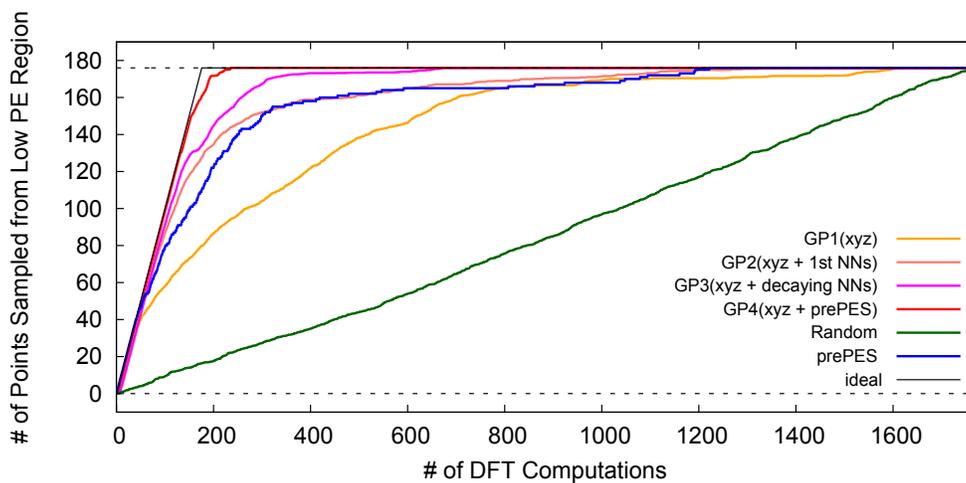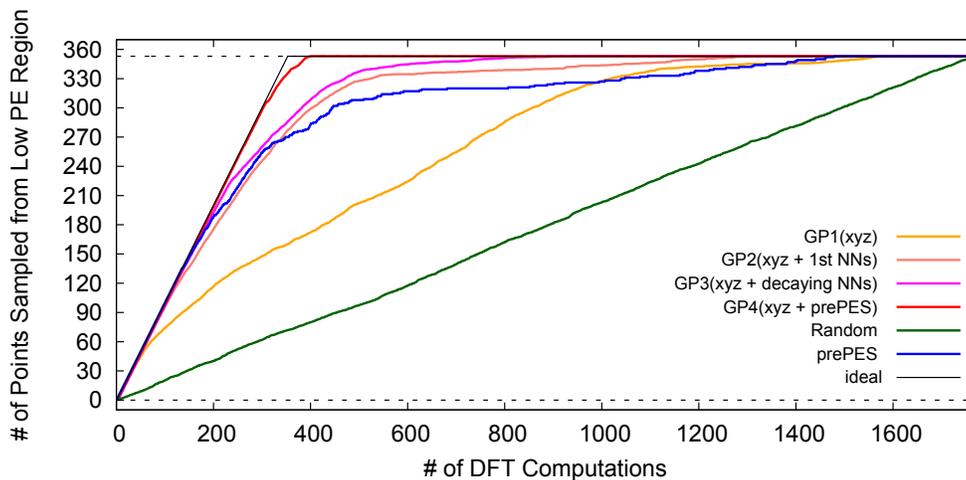
FIG. 4. Efficiencies of the seven sampling methods for (a) $\alpha = 0.05$, (b) $\alpha = 0.10$ and (c) $\alpha = 0.20$. The numbers of grid points successfully sampled from the low PE region (#(TP)) are plotted as functions of the number of PE evaluations by DFT computations (#(TP) + #(FP)).
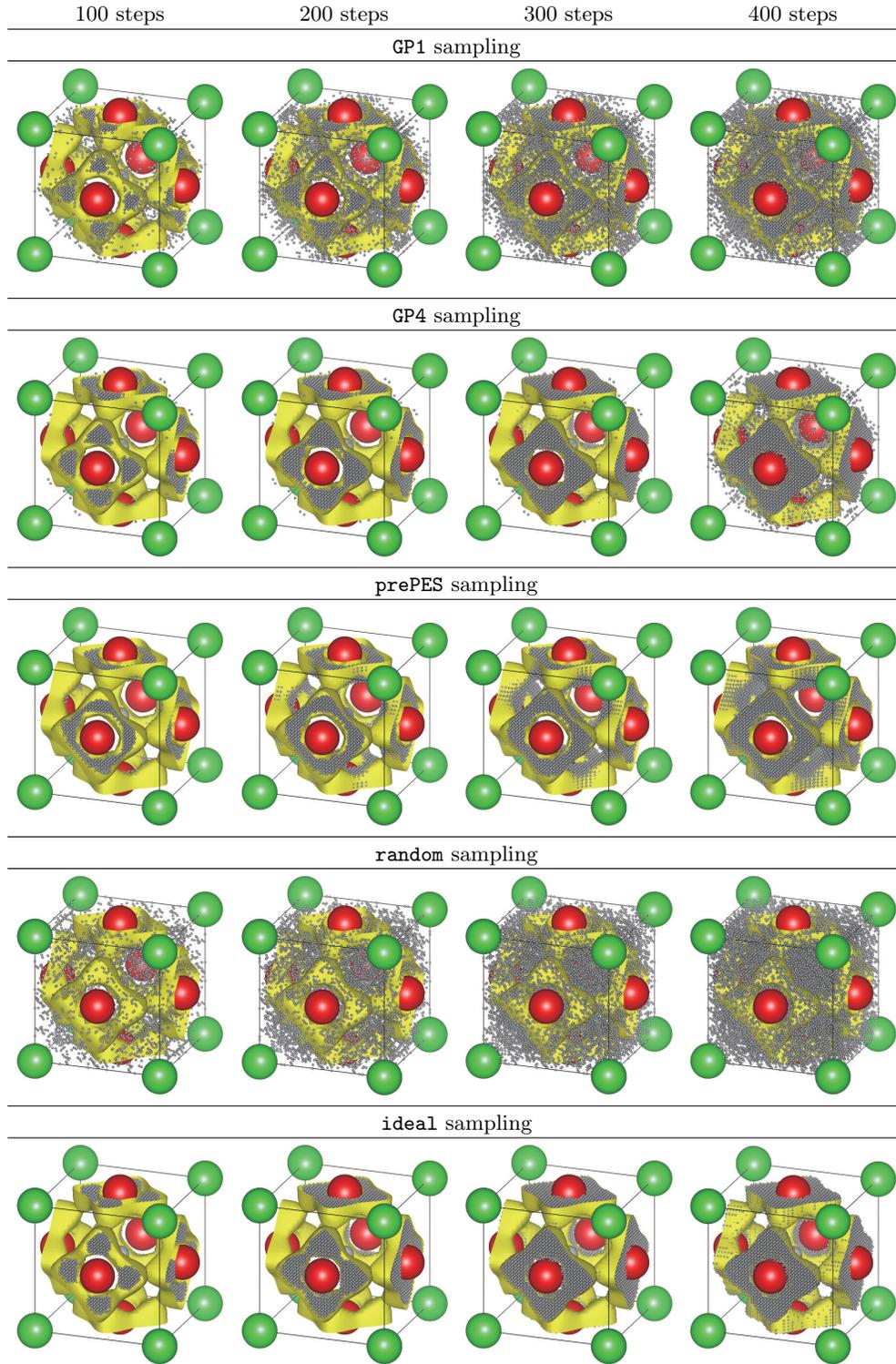
FIG. 5. Selected grid points (gray dots) at 100, 200, 300, and 400 steps by different sampling methods in the model crystal lattice of BaZrO₃ for $\alpha = 0.20$ case. The yellow surface in each plot is the isosurface corresponding to the energy threshold at $\alpha = 0.2$.
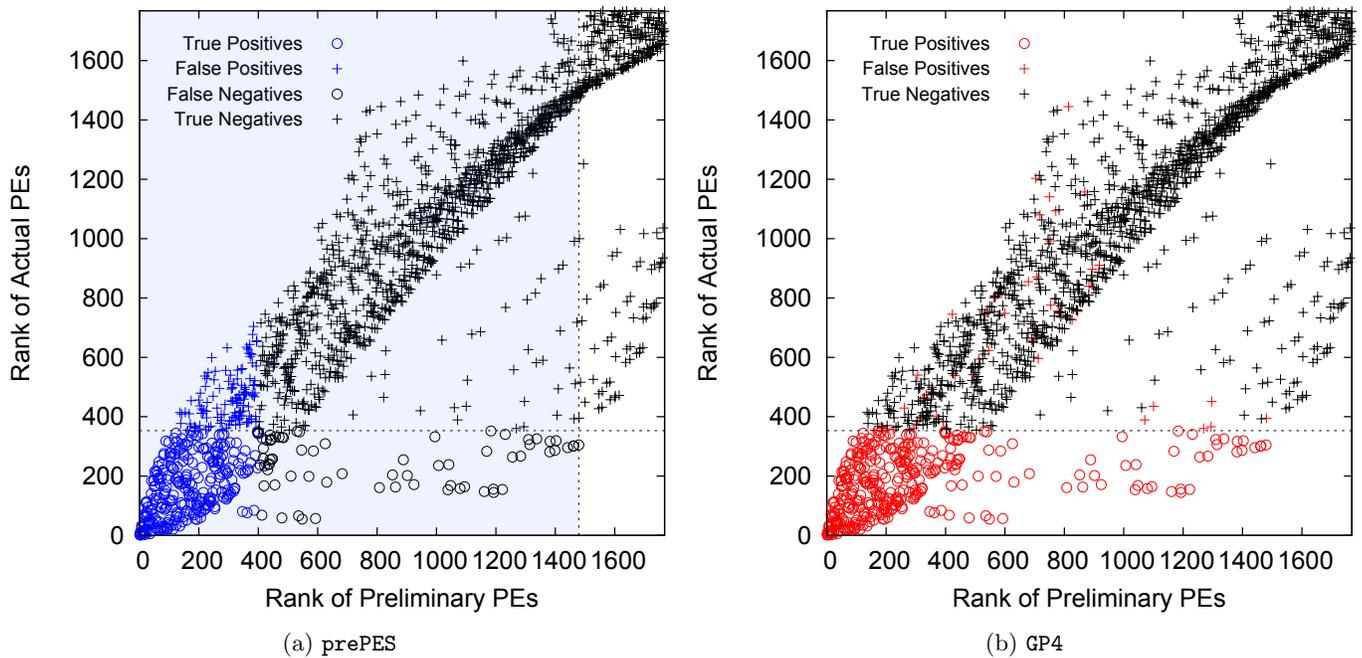
FIG. 6. Rank correlation between the actual and the preliminary PEs. The grid points in $\mathcal{P}_\alpha$ are shown as open circles, while those in $\mathcal{N}_\alpha$ are shown as crosses. Sampled points at 400 steps are also shown as blue and red symbols in (a) `prePES` and (b) `GP4`, respectively, for the $\alpha = 0.20$ case. The method `GP4` samples all the positive points at 400 steps with a small number of false positives, i.e., sampled points that are not included in the low PE region. Therefore, in (b), there are no false negative points.

FIG.7 shows the points that could not be found by the `prePES` sampling method. This reflects the displacement of the first-nearest-neighbor O ion (O_{1NN}) for keeping a suitable distance from the proton (~ 1 Å) in the case of ordinary DFT calculation with structural relaxation. In addition, the structural relaxation around a proton critically affects the potential barrier heights for the proton migration. The potential barrier of the rotational and hopping paths obtained by the single-point DFT calculations are 0.37 and 1.29 eV, respectively, which are much higher than the ground truth values obtained by the ordinary fully relaxed DFT calculations.

On the other hand, by using GP with `xyz` and `prePES` as descriptors, the points close to O atom were also successfully identified. This is due to the advantage of the GP-based procedure where we could successfully predict by using the GP model that the actual PE values and the preliminary PE values are highly different in the neighborhoods of O atoms. In `GP4`, the average number of sampling steps required for identifying all the points in the low PE region was only 394.1 which is only 1.1-fold of the `ideal` sampling method.

V. PRACTICAL ISSUES

In this section, we present results regarding two practical issues. In §V A, we show results on when the sam-

pling should be stopped. In §V B, we discuss and present results on how an appropriate α can be selected.

A. Stopping criterion

One of the important practical issues in selective sampling methods is when the sampling should be stopped. As we discussed in §II D, one of the practical advantages of statistical models such as the GP model is that we can estimate how many points still remain to be sampled by estimating the false negative rate (FNR). FIG.8 shows the profiles of the estimated FNR and the estimated threshold as functions of the number of DFT calculations in the case of $\alpha = 0.20$. These plots indicate that the estimated FNR almost coincides with the ground truth line, and that the estimated threshold also converges to the true value as the sampling proceeds. These results suggest that the estimated FNR would be actually useful as a stopping criterion.

B. How to determine an appropriate α value

Another practical issue is how to choose an appropriate α depending on what physical property is being investigated. In the case of proton conduction in an oxides, we should identify the low PE region in such a way that a
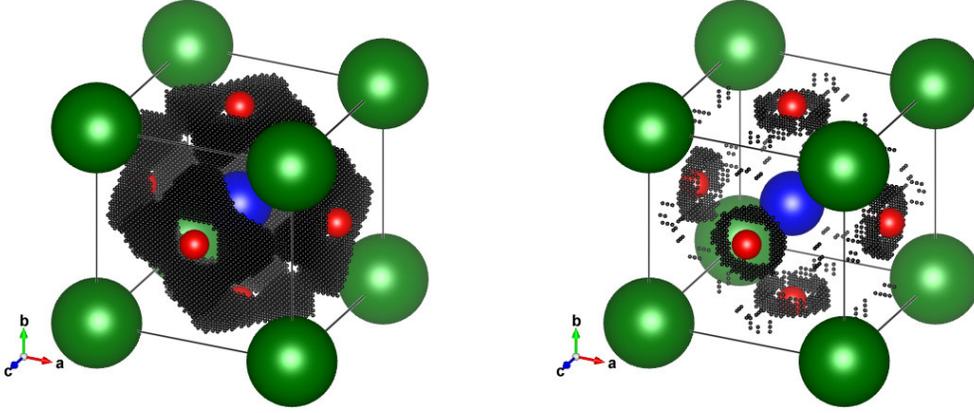
(a) Grid points in the low PE region with $\alpha = 0.2$ (b) Grid points that could not be found by `prePES`

FIG. 7. (a) Grid points in the low PE region with $\alpha = 0.2$ of $BaZrO_3$, and (b) grid points that could not be found via the `prePES` sampling method in 394 steps by which the `GP4` sampling method could successfully identify all the grid points in the low PE region. Plot (b) indicates that grid points around the O atom (red) could not be correctly identified by using preliminary PES alone, i.e., the preliminary PEs at these points are NOT relatively low, although their actual PEs are found to be relatively low after structural optimization.
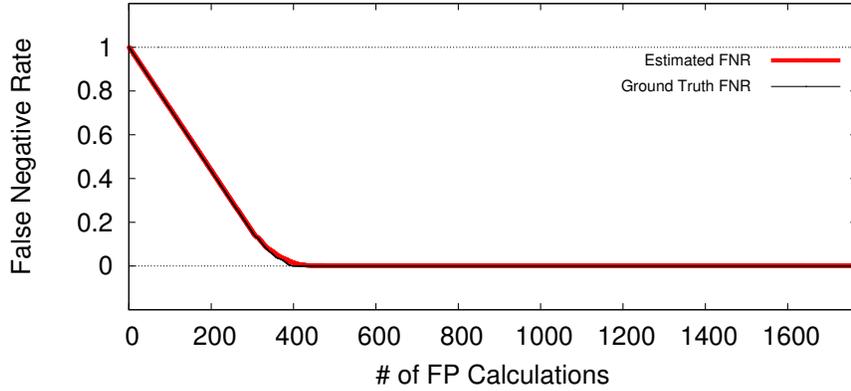
(a) The estimated FNR
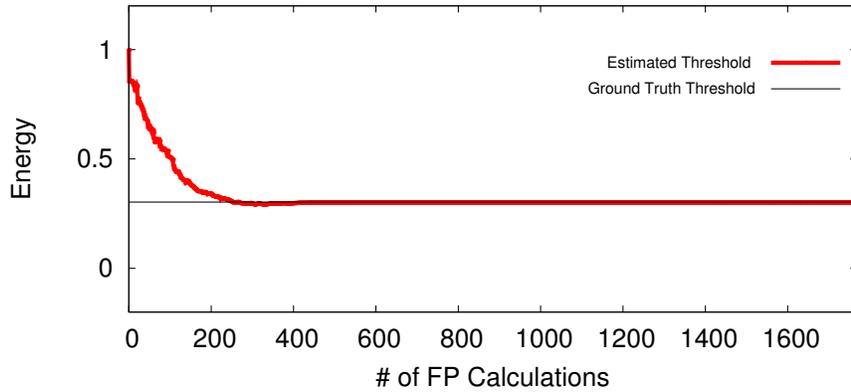
(b) The estimated threshold

FIG. 8. The profiles of the estimated FNRs and thresholds for $\alpha = 0.20$ case.

(a) Sampling profile

(b) FNR
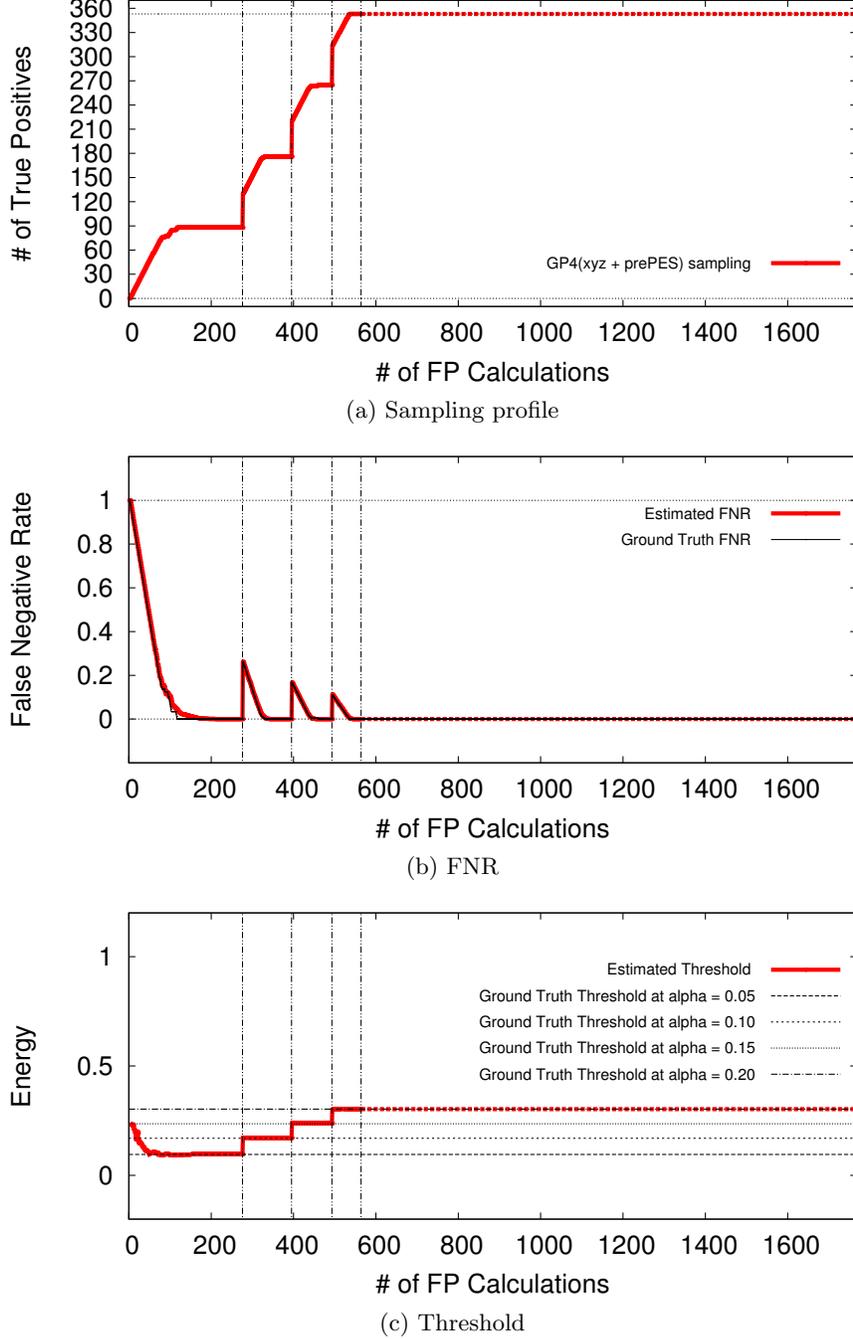
(c) Threshold

FIG. 9. Results when α is gradually increased as $0.05 \rightarrow 0.10 \rightarrow 0.15 \rightarrow 0.20$ in a stepwise manner. (a) THe true positive points, (b) the estimated FNR, and (c) the estimated threshold are plotted as functions of the numbers of DFT computations. Here, α is updated when the estimated false negative rate $\hat{\text{FNR}}$ becomes less than $\varepsilon = 10^{-6}$ at the current α. Note that the FNR in (b) and the threshold in (c) change with α.

proton-conducting network exists throughout the crystal lattice within the region. According to the actual PEs in FIG.3, the low PE regions are isolated when $\alpha < 0.15$, while they are abruptly connected when $\alpha \simeq 0.20$. This means that a proper α value would be around 0.20 in the present study.

If such an appropriate α value is unknown beforehand, we can set the α value in a stepwise manner. For demonstrating this approach, we investigated the performance of GP4 when α was increased as 0.05, 0.10, 0.15 and 0.20 in a stepwise manner, whose results are shown in FIG.9. In this scenario, we increased α by 0.05 when the es-

timated FNR became smaller than 10^{-6}. The plot (b) indicates that, in the first step with $\alpha = 0.05$, the convergence of the estimated FNR was a bit slower than the ground truth FNR. This is why we had to sample more than 250 times before we were sure that all the points in $\mathcal{P}_{0.05}$ had been successfully sampled. On the other hand, when $\alpha = 0.10$, 0.15, and 0.20, the convergences of the FNRs were almost as fast as the ground truth FNRs. Note that the true positive points abruptly increase when the α value is switched, indicating that the positive points for higher α have already been sampled in earlier steps. Although this stepwise strategy was a little less efficient than directly specifying $\alpha = 0.20$, it was still much more efficient than the `prePES` and `random` sampling methods.

VI. CONCLUSION

We introduced a machine learning-based selective sampling procedure for PES evaluation, and demonstrated the efficacy of the proposed procedure by applying it to proton conduction in $BaZrO_3$. The performance of the selective sampling method based on the GP model greatly depends on the descriptors, and the use of a preliminary PES (`GP4(xyz + prePES)`) is significantly effective, which was here evaluated by single-point DFT calculations in a smaller supercell. This indicates that the machine learning approach hybridized with a low-cost PES evaluation is a solid methodology for preferential PES evaluation in a region of interest. In addition, we demonstrated that two practical issues, namely, when to stop the sampling and how to determine an appropriate α value (equivalent to the PE threshold), can be solved by using the false negative rate (FNR) defined in Eq. (12).

[1] S. J. Lombardo and A. T. Bell, Surface science reports **13**, 3 (1991).

[2] F. W. Starr, S. Sastry, E. La Nave, A. Scala, H. E. Stanley, and F. Sciortino, Physical Review E **63**, 041201 (2001).

[3] A. Fernández-Ramos, J. A. Miller, S. J. Klippenstein, and D. G. Truhlar, Chemical reviews **106**, 4518 (2006).

[4] X. Huang, B. J. Braams, J. M. Bowman, R. E. Kelly, J. Tennyson, G. C. Groenenboom, and A. van der Avoird, The Journal of chemical physics **128**, 034312 (2008).

[5] O. Hellman, I. Abrikosov, and S. Simak, Physical Review B **84**, 180301 (2011).

[6] J. C. Thomas and A. Van der Ven, Physical Review B **88**, 214111 (2013).

[7] K. Reuter and M. Scheffler, Physical Review B **73**, 045433 (2006).

[8] M. Corral Valero, P. Raybaud, and P. Sautet, The Journal of Physical Chemistry B **110**, 1759 (2006).

[9] M. Rosini, R. Magri, and P. Kratzer, Physical Review B **77**, 165323 (2008).

[10] G. H. Vineyard, Journal of Physics and Chemistry of Solids **3**, 121 (1957).

[11] G. Henkelman, B. P. Uberuaga, and H. Jónsson, The Journal of chemical physics **113**, 9901 (2000).

[12] M. E. Björketun, P. G. Sundell, G. Wahnström, and D. Engberg, Solid State Ionics **176**, 3035 (2005).

[13] H. Iwahara, T. Yajima, T. Hibino, K. Ozaki, and H. Suzuki, Solid State Ionics **61**, 65 (1993).

[14] W. Münch, K.-D. Kreuer, G. Seifert, and J. Maier, Solid State Ionics **136**, 183 (2000).

[15] M. S. Islam, Journal of Materials Chemistry **10**, 1027 (2000).

[16] T. Gu, W. Lu, X. Bao, and N. Chen, Solid state sciences **8**, 129 (2006).

[17] J. Behler and M. Parrinello, Physical review letters **98**, 146401 (2007).

[18] L. Xu, L. Wencong, P. Chunrong, S. Qiang, and G. Jin, Computational Materials Science **46**, 860 (2009).

[19] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Physical review letters **104**, 136403 (2010).

[20] W. Setyawan, R. M. Gaume, S. Lam, R. S. Feigelson, and S. Curtarolo, ACS combinatorial science **13**, 382 (2011).

[21] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Physical review letters **108**, 058301 (2012).

[22] Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Physical Review B **85**, 104104 (2012).

[23] K. Fujimura, A. Seko, Y. Koyama, A. Kuwabara, I. Kishida, K. Shitara, C. A. Fisher, H. Moriwake, and I. Tanaka, Advanced Energy Materials **3**, 980 (2013).

[24] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, New Journal of Physics **15**, 095003 (2013).

[25] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, Physical Review B **89**, 054303 (2014).

[26] A. Seko, A. Takahashi, and I. Tanaka, Physical Review B **90**, 024101 (2014).

[27] K. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. Müller, and E. Gross, Physical Review B **89**, 205118 (2014).

[28] P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon, and K. Rajan, Computational Materials Science **83**, 185 (2014).

[29] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Physical review letters **114**, 105503 (2015).

[30] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, International Journal of Quantum Chemistry **115**, 1094 (2015).

[31] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, Physical Review B **89**, 094104 (2014).

[32] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, arXiv preprint arXiv:1508.05315 (2015).

[33] R. Ramakrishnan, P. O. Dral, M. Rupp, and A. O. von Lilienfeld, Journal of Chemical Theory and Computation **11**, 2087 (2015).

[34] C. K. Williams and C. E. Rasmussen, the MIT Press **2**, 4 (2006).

[35] M. L. Stein, Interpolation of spatial data: some theory for kriging (Springer Science & Business Media, 2012).

[36] J. Mockus, Journal of Global Optimization **4**, 347 (1994).

[37] E. Brochu, V. M. Cora, and N. De Freitas, arXiv preprint arXiv:1012.2599 (2010).

[38] T. Pagnier, I. Charrier-Cougoulic, C. Ritter, and G. Lucazeau, The European Physical Journal Applied Physics **9**, 1 (2000).

[39] I. Levin, T. G. Amos, S. M. Bell, L. Farber, T. A. Vanderah, R. S. Roth, and B. H. Toby, Journal of Solid State Chemistry **175**, 170 (2003).

[40] P. E. Blöchl, Physical Review B **50**, 17953 (1994).

[41] G. Kresse and J. Hafner, Physical Review B **48**, 13115 (1993).

[42] G. Kresse and J. Furthmüller, Computational Materials Science **6**, 15 (1996).

[43] G. Kresse and D. Joubert, Physical Review B **59**, 1758 (1999).

[44] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical review letters **77**, 3865 (1996).

[45] M. A. Gomez, M. A. Griffin, S. Jindal, K. D. Rule, and V. R. Cooper, The Journal of chemical physics **123**, 094703 (2005).

[46] N. Bork, N. Bonanos, J. Rossmeisl, and T. Vegge, Physical Review B **82**, 014103 (2010).