Quantification of uncertainties in thermoelectric properties of materials from a first-principles prediction method: An approach based on Gaussian process regression

Daehyun Wee, Jeeyoung Kim, Semi Bang, Georgy Samsonidze, and Boris Kozinsky

Phys. Rev. Materials **3**, 033803 — Published 25 March 2019

DOI: 10.1103/PhysRevMaterials.3.033803

Quantification of Uncertainties in Thermoelectric Properties of Materials from a First-Principles Prediction Method: An Approach Based on Gaussian Process Regression

Daehyun Wee,* Jeeyoung Kim, and Semi Bang

Department of Environmental Science and Engineering,

Ewha Womans University, Seoul, 03760, Republic of Korea

Georgy Samsonidze

Research and Technology Center, Robert Bosch LLC, Cambridge, MA 02142, U.S.A.

Boris Kozinsky

Harvard John A. Paulson School of Engineering and Applied Sciences,

Harvard University, Cambridge, MA 02138, U.S.A.

(Dated: February 27, 2019)

Abstract

We present the electron-phonon averaged via Gaussian process regression (EPA-GPR) method, in which the electron-phonon coupling matrix is represented as a function of two energies and is in turn modeled as a Gaussian process. The EPA-GPR method can be used as an efficient method to estimate thermoelectric properties of materials for fast-screening applications, comparable to the original electron-phonon averaged (EPA) method and the electron-phonon averaged via moving-least-squares (EPA-MLS) method. The EPA-GPR method does not require specification of any open parameter, unlike the other EPA-related methods, since all the hyperparameters in the model can be unambiguously estimated within the type II maximum likelihood (ML-II) approximation. Thus, the EPA-GPR method is a parameter-free estimation method. Additionally, the concept of Gaussian processes in the EPA-GPR method allows us to quantify the uncertainty in estimated properties of thermoelectric materials. One can randomly realize the electron-phonon coupling coefficients from the identified Gaussian process, and those realized samples can be further analyzed in the solution process of the semiclassical Boltzmann transport equation for charge carriers. The results of the semiclassical Boltzmann transport equation provide the statistical properties of the thermoelectric properties of interest. The means, standard deviations, histograms, and confidence intervals of the Seebeck coefficient, the electrical conductivity, and the power factor can be constructed and analyzed. The proposed EPA-GPR method is applied to a p-type half-Heusler compound, i.e., HfCoSb, as a case example, the results of which clearly present the advantages of the method.

I. INTRODUCTION

The performance of thermoelectric (TE) energy conversion, where thermal energy is directly converted into electrical energy or vice versa by using thermoelectric materials, depends on the thermoelectric figure of merit, i.e., ZT of the material[1]. ZT is defined as a combination of thermal and electrical properties of the material, as follows:

$$ZT = \frac{S^2 \sigma T}{k}, \tag{1}$$

where S is the Seebeck coefficient, σ is the electrical conductivity, k is the thermal conductivity of the material, and T is the operating temperature. Researchers in the computational materials science community interested in thermoelectric energy conversion have exerted tremendous efforts to develop methods that can be used for estimating these properties from the first principles to discover better thermoelectric materials. For example, many computational studies estimating k and other related thermal properties have been reported[2–4].

On the other hand, the electrical properties of inorganic materials, including S and σ, can be obtained by solving the semiclassical Boltzmann transport equation within the relaxation time approximation[5]. The simplest approach that can be used for estimation of these electronic transport properties is the constant relaxation time (CRT) approximation, in which one single value for relaxation time τ is arbitrarily assumed[6]. However, such an approach naturally introduces an arbitrary constant, i.e., relaxation time, and does not possess any predictive capacity, rendering the approach unsatisfactory for screening thermoelectric materials from the first principles. It is, therefore, necessary to develop a more predictive method for dealing with the relaxation time.

Matthiessen's rule states that the total scattering rate τ^{-1} of electrons is the sum of the rates associated with intrinsic (electron-electron, electron-phonon) and extrinsic (impurities, grain boundaries, alloy disorder) scattering mechanisms. To screen potentially promising candidates for thermoelectric applications, one must first identify the intrinsic properties of the material, since the extrinsic properties are tuned during the synthesis process. In automotive TE power generation, the relevant temperature is around 400°C at the hot side of the device, at which electron-phonon (el-ph) interaction becomes the dominant scattering mechanism[7,8]. The first-principles estimation of the el-ph interaction has been pursued

by several different approaches with various levels of computational complication. The deformation potential (DP) approximation is one of the simplest approaches[9], but such simplification often cannot be justified for complex TE materials. The other extreme is the electron-phonon Wannier (EPW) method[10], which fully describes the el-ph scattering. However, the EPW method is not appropriate for fast-screening applications due to its high computational cost.

More recently, a new approach, i.e., the electron-phonon averaged (EPA) method, which combines simplicity and speed with a fully first-principles treatment of the el-ph interaction, has been introduced[11]. By turning the complex momentum-space integration into an integration over energies and simultaneously replacing several terms with their averages within bins over an energy range, the EPA method allows for automated rapid calculations for optimization of electronic transport quantities, while being more predictive than the CRT and DP approximations. The method has been successfully used for screening potential TE materials from a group of half-Heusler (HH) compounds[11]. Later, it was proposed to modify the standard EPA method through combination with a moving-least-squares (MLS) averaging strategy[12]. It was demonstrated that the electron-phonon averaged via moving-least-squares (EPA-MLS) method could make a similar prediction of thermoelectric properties of materials with a much coarser momentum grid than was required for the standard EPA method[12].

However, several limitations remain in the EPA and EPA-MLS methods. First, these methods require specification of an open parameter, i.e., either the bin size or the length scale of the smoothing kernel. Second, although the sample variance during the averaging process can roughly provide the amount of uncertainty in the estimated electron-phonon coupling effects, rigorous analysis of uncertainty and sensitivity can be difficult within the EPA and EPA-MLS methods. The first problem is of minor importance, especially because the result of the EPA-MLS method seems rather insensitive to particular choices of the open parameter[12]. The lack of a rigorous strategy for uncertainty quantification in the numerical procedure is a more serious issue that requires immediate attention. One should not place blind confidence in his or her prediction without describing the underlying uncertainty. The same issue is essentially shared by most of the first-principles methods used in the study of thermoelectric properties. None of the methods we have mentioned so far, i.e., the CRT, DP, and EPW methods, currently has a rigorous quantification strategy of uncertainty in its numerical procedures.

In this paper, we investigate the possibility of using a mathematically rigorous alternative method. Here, the electron-phonon coupling matrix is modeled as a Gaussian process, which is widely used in the context of machine learning[13]. During regression, the characteristic length-scale of the covariance function of the Gaussian process, which serves a similar purpose to the smoothing scale in the EPA-MLS method, can be estimated within the type II maximum likelihood (ML-II) approximation without any ambiguity. At the same time, all the statistical tools that can be used for Gaussian processes are readily available for further analysis of uncertainty and sensitivity of the results. The resulting formulation, i.e., the electron-phonon averaged via Gaussian process regression (EPA-GPR) method, may resolve the above issues.

The paper is organized as follows. We first describe the basic theory of the EPA-GPR method in Section II. We continue to test the method on a p-type HH compound in Section III. The values of the thermoelectric properties, i.e., S and σ, estimated by the EPA-GPR method are compared to those using other related methods and experiments. The uncertainties in the thermoelectric properties are also quantified by the method described in Section II. A brief summary follows in Section IV.

II. THEORY

A. The EPA and EPA-MLS Methods

We first briefly review the main features of the EPA and EPA-MLS methods. Details may be found in[12]. The main task of predicting the electronic transport coefficients for electrons within the relaxation time approximation is evaluation of the inverse of the electron energy relaxation time induced by the electron-phonon (el-ph) interaction, which is given as follows[14,15]:

$$\begin{aligned}\tau_{n\mathbf{k}}^{-1}(\mu,T) = \frac{\Omega}{(2\pi)^2\,\hbar}\sum_{m\nu}\int_{BZ} d\mathbf{q}\left|g^{\mathrm{SE}}_{mn\nu}(\mathbf{k},\mathbf{q})\right|^2 \\ \times\Big\{\Big[n(\omega_{\nu\mathbf{q}},T)+f(\epsilon_{m\mathbf{k}+\mathbf{q}},\mu,T)\Big]\delta\left(\epsilon_{n\mathbf{k}}+\omega_{\nu\mathbf{q}}-\epsilon_{m\mathbf{k}+\mathbf{q}}\right) \\ +\Big[n(\omega_{\nu\mathbf{q}},T)+1-f(\epsilon_{m\mathbf{k}+\mathbf{q}},\mu,T)\Big]\delta\left(\epsilon_{n\mathbf{k}}-\omega_{\nu\mathbf{q}}-\epsilon_{m\mathbf{k}+\mathbf{q}}\right)\Big\}, \quad (2)\end{aligned}$$

where Ω is the volume of the primitive cell, m and n are the electron band indices, ν is the phonon mode index, $\mathbf{k}$ is the electron wavevector, $\mathbf{q}$ is the phonon wavevector, $\epsilon_{n\mathbf{k}}$ is the electron energy, $\omega_{\nu\mathbf{q}}$ is the phonon energy, $g^{\mathrm{SE}}_{mn\nu}(\mathbf{k},\mathbf{q})$ is the el-ph coupling matrix element, $n(\omega,T)$ is the Bose-Einstein distribution function, $f(\epsilon,\mu,T)$ is the Fermi-Dirac distribution function, δ is the Dirac delta function, μ is the chemical potential of electrons, k_B is the Boltzmann constant, and $\hbar$ is the reduced Planck constant.

The el-ph coupling matrix elements, i.e., $g^{\mathrm{SE}}_{mn\nu}(\mathbf{k},\mathbf{q})$, can be obtained from the DFPT calculations[16], which are relatively costly for materials with a large unit cell. The main element of the EPA approximation is to replace the momentum-dependent quantities in Eq. (2) with their energy-dependent averages. Accordingly, the el-ph coupling matrix elements are averaged over the directions of $\mathbf{k}$ and $\mathbf{k}+\mathbf{q}$ wavevectors:

$$\left|g^{\mathrm{SE}}_{mn\nu}(\mathbf{k},\mathbf{q})\right|^2 \mapsto g^2_\nu(\epsilon_{n\mathbf{k}},\epsilon_{m\mathbf{k}+\mathbf{q}}). \tag{3}$$

As a result, g^2_ν becomes a function of two energies, ϵ_1 and ϵ_2, which represent the energy of the incoming electron state and that of the outgoing electron state, respectively. Additionally, $\omega_{\nu\mathbf{q}}$ is also replaced with its average:

$$\omega_{\nu\mathbf{q}} \mapsto \bar{\omega}_\nu. \tag{4}$$

With these substitutions, the integration over $\mathbf{q}$ and the summation over m in Eq. (2) can be evaluated analytically, yielding

$$\begin{aligned}\tau^{-1}(\epsilon,\mu,T) = \frac{2\pi\Omega}{g_s\,\hbar}\sum_\nu \\ \Big\{ g^2_\nu(\epsilon,\epsilon+\bar{\omega}_\nu)\Big[n(\bar{\omega}_\nu,T)+f(\epsilon+\bar{\omega}_\nu,\mu,T)\Big]\rho\left(\epsilon+\bar{\omega}_\nu\right) \\ + g^2_\nu(\epsilon,\epsilon-\bar{\omega}_\nu)\Big[n(\bar{\omega}_\nu,T)+1-f(\epsilon-\bar{\omega}_\nu,\mu,T)\Big]\rho\left(\epsilon-\bar{\omega}_\nu\right)\Big\}.\end{aligned} \tag{5}$$

Here, $\rho(\epsilon)$ is the electron density of states defined as the number of electronic states per unit energy and unit volume, and $g_s = 2$ is the spin degeneracy.

Various methods can be used to achieve the mapping of Eq. (3). In the original EPA method[11], a bin-based averaging strategy was employed with a predefined bin size δ_{Bin}. On

the other hand, in[12], we proposed the use of an MLS averaging strategy[17], where $g^2_\nu(\epsilon_1, \epsilon_2)$ for each pair of ϵ_1 and ϵ_2 is obtained by minimizing

$$\sum_{mn} \iint_{BZ} d\mathbf{k} d\mathbf{q} \left(g^2_\nu(\epsilon_1, \epsilon_2) - \left| g^{\mathrm{SE}}_{mn\nu}(\mathbf{k}, \mathbf{q}) \right|^2 \right)^2 \times \exp\left(- \frac{(\epsilon_{n\mathbf{k}} - \epsilon_1)^2 + (\epsilon_{m\mathbf{k}+\mathbf{q}} - \epsilon_2)^2}{2\sigma^2_{\mathrm{Gauss}}} \right), \quad (6)$$

in which σ_{Gauss} represents the smoothing scale of the Gaussian function. Since BZ integrations are typically performed by sampling over $\mathbf{k}$ and $\mathbf{q}$-point grids, the expression for $g^2_\nu(\epsilon_1, \epsilon_2)$ is given by the weighted sample mean of $\left| g^{\mathrm{SE}}_{mn\nu}(\mathbf{k}, \mathbf{q}) \right|^2$. Setting $V_1 = \sum_{mn\mathbf{kq}} w_{mn\mathbf{kq}}$, where $w_{mn\mathbf{kq}}$ is the weight of each sample, including both the degeneracy of the sample point in the Brillouin zone and the Gaussian factor shown in Eq. (6), we get

$$g^2_\nu(\epsilon_1, \epsilon_2) = \frac{1}{V_1} \sum_{mn\mathbf{kq}} w_{mn\mathbf{kq}} \left| g^{\mathrm{SE}}_{mn\nu}(\mathbf{k}, \mathbf{q}) \right|^2. \quad (7)$$

Since the phonon calculations typically dominate the computational cost during the entire calculation process during the evaluation of electron-phonon coupling matrix, the use of a coarser $\mathbf{q}$-point grid directly leads to an almost proportional reduction of the overall computational cost. It was reported in[11] that the el-ph calculation, i.e., the phonon calculation, took about 100 core-hours on 4×4×4 $\mathbf{q}$-point grids and 4600 core-hours on 8×8×8 grids for a single HH compound. It was also reported that, for a given chemical potential of electrons and temperature, the CRT and EPA calculations took about 0.15 core-hours each when using 8×8×8 $\mathbf{q}$-point grids for phonon calculations and 48×48×48 $\mathbf{k}$-point grids for the band structure calculations, while a comparable EPW calculation took about 2600 core-hours when using 4×4×4 and 32×32×32 grids[11]. On the other hand, it was shown in[12] that the EPA-MLS method could allow the use of a much coarser grid, i.e., 2×2×2 $\mathbf{q}$-point grid, for the phonon calculation with an acceptable result for fast-screening purposes.

While the use of the EPA-MLS method achieved a reasonable balance between performance and accuracy[12], there are still problems. First, the method still requires specification of an open parameter, the smoothing scale of the smoothing kernel. Although the computed results were not very sensitive to this parameter, it is an annoying nuisance. Second, although the sample variance can be used for a rough estimate of uncertainty in the estimated el-ph coupling effects[12], a rigorous analysis of uncertainty and sensitivity is rather difficult within the EPA-MLS method.

B. Gaussian Process Regression of Electron-Phonon Coupling

A more rigorous alternative method to achieve the transformation of Eq. (3) is to model $g_\nu^2(\epsilon_1, \epsilon_2)$ as a Gaussian process[13] and to perform regression based on the observed elements of the electron-phonon coupling matrix. Then, during Gaussian process regression (GPR), the characteristic length-scale of the covariance function of the Gaussian process, which serves the same purpose as the smoothing scale in the EPA-MLS method, can be estimated using the type II maximum likelihood (ML-II) approximation. At the same time, the analysis of uncertainty and sensitivity can become theoretically more straightforward.

Formally, a Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution[13]. In this paper, g_ν^2 is modeled as a Gaussian process. Thus,

$$g_\nu^2(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{8}$$

where $\mathbf{x}$ represents the two-dimensional vector coordinate (ϵ_1, ϵ_2), and $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ are the mean and covariance functions of $g_\nu^2(\mathbf{x})$, respectively. We consider 0 as the mean, since virtually no prior knowledge is available. Many different covariance functions can be used, but a simple square exponential kernel is employed here as the covariance function of the choice:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_{\mathrm{SEK}}^2 \exp\left(-\frac{(\epsilon_1 - \epsilon_1')^2 + (\epsilon_2 - \epsilon_2')^2}{2\ell_{\mathrm{SEK}}^2}\right). \tag{9}$$

ℓ_{SEK} is the correlation length scale of the Gaussian process, essentially playing the same role as that of σ_{Gauss} in the EPA-MLS method.

Then, we make noisy observations of g_ν^2 at various training points in two-dimensional energy space. The set of training points is denoted as X, and the DFPT calculations of the values of $\left|g_{mn\nu}^{\mathrm{SE}}(\mathbf{k}, \mathbf{q})\right|^2$ on these training points are considered to be such observations. That is,

$$\left|g_{mn\nu}^{\mathrm{SE}}(\mathbf{k}, \mathbf{q})\right|^2 = g_\nu^2(\epsilon_{n\mathbf{k}}, \epsilon_{m\mathbf{k}+\mathbf{q}}) + \epsilon_{\mathrm{noise}}, \tag{10}$$

where $\epsilon_{\mathrm{noise}}$ is additive, independent, identically distributed Gaussian noise with variance $\sigma_{\mathrm{noise}}^2$. The total number of observed $\left|g_{mn\nu}^{\mathrm{SE}}(\mathbf{k}, \mathbf{q})\right|^2$ is N_S.

The objective of Gaussian process regression is to predict the values of g_ν^2 at N_T test points $\mathbf{x}_{*,j}$ $(1 \leq j \leq N_T)$. The training vector is constructed by combining the noisy observations, i.e., $\mathbf{y} = \left[\left| g_{mn\nu}^{\mathrm{SE}} \right|_1^2 \ \left| g_{mn\nu}^{\mathrm{SE}} \right|_2^2 \ \left| g_{mn\nu}^{\mathrm{SE}} \right|_3^2 \cdots \left| g_{mn\nu}^{\mathrm{SE}} \right|_{N_S}^2 \right]^\top$. We define the test output vector as $\mathbf{f}_* = \left[g_{\nu*,1}^2 \ g_{\nu*,2}^2 \ g_{\nu*,3}^2 \cdots g_{\nu*,N_T}^2 \right]^\top$, where $g_{\nu*,j}^2$ is the estimated value of $g_\nu^2(\mathbf{x}_{*,j})$ plus an additive noise:

$$g_{\nu*,j}^2 = g_\nu^2(\mathbf{x}_{*,j}) + \epsilon_{\mathrm{noise}}. \tag{11}$$

Here, $\epsilon_{\mathrm{noise}}$ has the same variance $\sigma_{\mathrm{noise}}^2$ as in Eq. (10). The definition of the test output vector in this paper is slightly different from that of typical GPR cases. Typically, the test output is specified as the estimated value of the Gaussian process only, excluding any additive noise. However, in our case, the Gaussian process, i.e., g_ν^2, is only an approximate representation of the physical quantity of interest, i.e., $\left| g_{mn\nu}^{\mathrm{SE}}(\mathbf{k}, \mathbf{q}) \right|^2$. Since $\left| g_{mn\nu}^{\mathrm{SE}}(\mathbf{k}, \mathbf{q}) \right|^2$ is the sum of g_ν^2 and $\epsilon_{\mathrm{noise}}$ as represented in Eq. (10), it is more appropriate to include $\epsilon_{\mathrm{noise}}$ during realization of the random Gaussian process, which must reproduce not g_ν^2 but $\left| g_{mn\nu}^{\mathrm{SE}}(\mathbf{k}, \mathbf{q}) \right|^2$.

According to the prior, the joint distribution of the training vector and the test output vector is given as follows:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B} \end{bmatrix} \right), \tag{12}$$

where $\mathbf{A} = \mathbf{K}_{X,X} + \sigma_{\mathrm{noise}}^2 \mathbf{I}$, $\mathbf{B} = \mathbf{K}_{X_*,X_*} + \sigma_{\mathrm{noise}}^2 \mathbf{I}$, and $\mathbf{C} = \mathbf{K}_{X,X_*}$. $\mathbf{K}_{X,X_*}$ denotes the $N_S \times N_T$ matrix of the covariances evaluated at all pairs of training and test points, and the other entries are defined in a similar way. $\mathbf{I}$ represents an identity matrix of an appropriate size.

Applying a standard argument for multivariate Gaussian distributions to this distribution[13], we construct the conditional distribution, which provides the key predictive equations for Gaussian process regression:

$$\mathbf{f}_* | X, \mathbf{y}, X_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \mathrm{cov}(\mathbf{f}_*)), \tag{13}$$

where

$$\bar{\mathbf{f}}_* = \mathrm{E}[\mathbf{f}_*|X, \mathbf{y}, X_*] = \mathbf{C}^\top \mathbf{A}^{-1} \mathbf{y}, \tag{14}$$

and

$$\mathrm{cov}(\mathbf{f}_*) = \mathbf{B} - \mathbf{C}^\top \mathbf{A}^{-1} \mathbf{C}. \tag{15}$$

The log marginal likelihood is given as follows:

$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^\top \mathbf{A}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{A}| - \frac{N_S}{2}\log 2. \tag{16}$$

To complete the specification of the model, we need to determine hyperparameters. There are three hyperparameters in the model: ℓ_{SEK}, σ_{SEK}, and σ_{noise}. One of the most widely used methods for identification of hyperparameters is the type II maximum likelihood (ML-II) approximation, in which the marginal likelihood of the available observations, i.e., Eq. (16), under the model is maximized with respect to the hyperparameters[13]. In this fashion, all the hyperparameters in the model, i.e., ℓ_{SEK}, σ_{SEK}, and σ_{noise}, are estimated.

The actual GPR procedure is performed using the Gaussian Processes for Machine Learning (GPML) Toolbox[18]. To reduce the computational cost, we use the KISS-GP (Kernel Interpolation for Scalable Structured Gaussian Processes) method[19], in which evaluation of the covariance function is replaced with interpolation from a well-defined grid. The calculation on the grid can exploit its underlying Kronecker-Toeplitz structure to boost the calculation speed, which renders the entire method practically feasible. The current implementation of the KISS-GP method requires us to use two separate correlation length scales, i.e., one for ϵ_1 and the other for ϵ_2. To recover one single length scale, the mean of these two lengths is calculated as ℓ_{SEK} after application of the ML-II approximation, which is the method that we use in this study.

C. Uncertainty Quantification of Thermoelectric Properties

One of the main advantages of the proposed GPR procedure is that it provides information on uncertainty. For example, the variance from Eq. (15) can be used as an indicator of the confidence interval for the estimated values of the el-ph coupling matrix elements at the test points. However, our main interest is not to quantify the uncertainty in the el-ph coupling

matrix elements but to quantify the uncertainty in the thermoelectric properties of the material itself, i.e., S and σ. For this purpose, we propose a simple Monte Carlo approach. Statistically, $\mathbf{f}_*$ follows a multivariate Gaussian distribution, whose mean and covariance are given by Eqs. (14-15). Thus, we can randomly realize $\mathbf{f}_*$ using the multivariate Gaussian statistic. Since $\mathbf{f}_*$ represents the estimated values of the el-ph coupling matrix elements, the semiclassical Boltzmann transport equation can be solved for each realization of $\mathbf{f}_*$ to create a sample of S and σ. By repeating this realization, one can create a large-sized set of realized samples, which will be further diagnosed to obtain the statistics of S and σ.

To reduce the computational cost further, we employ the following approximation:

$$g_\nu^2(\epsilon_1, \epsilon_2) \approx g_\nu^2 \left(\frac{\epsilon_1 + \epsilon_2}{2}, \frac{\epsilon_1 + \epsilon_2}{2} \right). \tag{17}$$

This is a valid approximation, since we only need $g_\nu^2(\epsilon, \epsilon \pm \bar{\omega}_\nu)$ for evaluating Eq. (5). The values of $\bar{\omega}_\nu$ are typically smaller than 0.1 eV, while the values of ℓ_{SEK} are about 1 eV. Since $\bar{\omega}_\nu \ll \ell_{\mathrm{SEK}}$,

$$g_\nu^2(\epsilon, \epsilon \pm \bar{\omega}_\nu) \approx g_\nu^2 \left(\epsilon \pm \frac{\bar{\omega}_\nu}{2}, \epsilon \pm \frac{\bar{\omega}_\nu}{2} \right), \tag{18}$$

because g_ν^2 will not vary much within the smoothing length scale of the Gaussian process, i.e., ℓ_{SEK}. The approximation allows us to use test points on the diagonal line ($\epsilon_1 = \epsilon_2$) only, which can be later extrapolated onto the two-dimensional energy space, using Eq. (17).

The overall procedure of uncertainty quantification within the electron-phonon averaged via Gaussian process regression (EPA-GPR) method can be summarized as follows.

1. Generate the training data set, i.e., $g_{mn\nu}^{\mathrm{SE}}(\mathbf{k}, \mathbf{q})$ on a coarse $\mathbf{q}$ mesh, from the DFPT calculations. For this purpose, we use the QUANTUM ESPRESSO package[20].

2. Perform the GPR procedure using the training data set. Fix the hyperparameters, i.e., ℓ_{SEK}, σ_{SEK}, and σ_{noise} by applying the ML-II approximation. All the operations in the GPR procedure are performed using the KISS-GP method in the GPML Toolbox[18].

3. Construct $\bar{\mathbf{f}}_*$ and $\mathrm{cov}(\mathbf{f}_*)$ for the test points on the diagonal line ($\epsilon_1 = \epsilon_2$) using Eqs. (14-15).

4. Randomly create N_R realized samples on the diagonal test points, using `mvnrnd`, which is a MATLAB function for random realization of the multivariate Gaussian statistic[21].

Extrapolate onto the two-dimensional energy space using Eq. (17). If an unrealistic negative value of g_ν^2 does occur, we put zero instead.

5. For each realized sample, construct the input files for a run of BoltzTraP[6]. BoltzTraP is a standard program for solving the semiclassical Boltzmann equation for inorganic semiconductors. The version used in this work is slightly modified from the original BoltzTraP program to incorporate an energy-dependent relaxation time.

6. Run BoltzTraP for N_R sample cases using the input files. Each BoltzTraP run is independent from the others. A large number of runs can be carried out in a parallel fashion within a relatively short time if enough computing power is available.

7. Statistically analyze the results of the BoltzTraP runs to quantify the uncertainty in S and σ.

III. NUMERICAL RESULTS

In this section, electronic transport properties for a TE material from the family of HH compounds, the p-type HfCoSb[22–24], are estimated to demonstrate the procedure explained in Section II. The HH compound has a MgAgAs structure type, whose space group is $F\bar{4}3m$[25,26]. After structural relaxation, the lattice parameter of the conventional cubic unit cell of the HfCoSb compound has a value around 6.0471 Å. The carrier concentration is fixed first at the value obtained from a Hall measurement at room temperature: $p = 0.06$ per formula unit (1.1×10^{21} cm^{-3}) for $Hf_{0.5}Zr_{0.5}CoSb_{0.8}Sn_{0.2}$[22].

The electron energy relaxation times and the electronic transport coefficients are calculated with the original EPA, EPA-MLS, and EPA-GPR methods. DFT and DFPT calculations are performed using the generalized gradient approximation in the PBE form[27] for exchange-correlation functional, ultrasoft pseudopotentials[28,29] for core-valence interaction and a plane wave basis set with 80 and 700 Ry kinetic energy cutoffs for wavefunctions and charge density. A uniform 8×8×8 Γ-centered **k**-point grid is used for self-consistent calculation of charge density, and 48×48×48 grids are used for band structure and transport calculations.

For the EPA method, a uniform 8×8×8 Γ-centered **q**-point grid is used for sampling $\left|g^{\mathrm{SE}}_{mn\nu}(\mathbf{k},\mathbf{q})\right|^2$ by direct el-ph calculations, which was the resolution used in a previous screen-

ing study[11]. For the EPA-MLS method, the same 8×8×8 Γ-centered **q**-point grid and a uniform 2×2×2 Γ-centered **q**-point grid are employed. For the EPA-GPR method, only the uniform 2×2×2 Γ-centered **q**-point grid is employed. Averaging in the EPA calculation is performed over the bins with $\delta_{\rm Bin} = 0.2$ eV—the smallest bin size at which all cells in the energy grid are filled with **k**-points. For the EPA-MLS method, $\sigma_{\rm Gauss} = 0.2$ eV. The hyperparameters in the EPA-GPR method are all identified within the ML-II approximation as described in Section II. To quantify the uncertainty in the EPA-GPR method, one thousand realized samples of g_ν^2 were generated and statistically analyzed.

First, the identified hyperparamters are presented in Table I. The values of $\ell_{\rm SEK}$ identified from the process are slightly less than 1 eV. Hence, the approximation of Eq. (17) can be considered valid. $\sigma_{\rm SEK}$ and $\sigma_{\rm noise}$ exhibit nontrivial sizes, suggesting that the averaging process of Eq. (3) involves significant uncertainty. The Seebeck coefficient and the electrical conductivity, computed from the original EPA, EPA-MLS, and EPA-GPR methods, are shown in Figure 1. We also plot the experimental data at similar conditions[22–24]. As mentioned earlier, the Hall measurement in[22] reported that the carrier concentration was 1.1×10^{21} cm^{-3} for $Hf_{0.5}Zr_{0.5}CoSb_{0.8}Sn_{0.2}$, which corresponds to $p = 0.06$ per formula unit. The doping concentration of the main dopant, i.e., Sn, for the nanostructured sample of $Hf_{0.8}Zr_{0.2}CoSb_{0.8}Sn_{0.2}$ in[23,24] was essentially the same to that for $Hf_{0.5}Zr_{0.5}CoSb_{0.8}Sn_{0.2}$ in[22], and hence it is expected that the nanostructured sample would have a similar carrier concentration, allowing us to compare our numerical results against those experimental data. It is clear that the EPA-GPR method shows good agreement to the other types of EPA method. In particular, the Seebeck coefficient is affected very little by the choice of a particular type of EPA method. On the other hand, the electrical conductivity shows slightly greater sensitivity to the choice. Also, our predicted values clearly exhibit reasonable correspondence to the experimental observations. Although the correspondence is not perfect, our predicted values maintain a level of accuracy that can be used in fast-screening applications.

Statistical properties of three different thermoelectric properties, i.e., the Seebeck coefficient (S), the electrical conductivity (σ), and the power factor (PF= $S^2\sigma$), at two different temperatures (300K and 700K) are summarized in Table II. The reference value ($a_{\rm ref}$) obtained from $\bar{\mathbf{f}}_*$ of Eq. (14), the mean ($\bar{a}$) and standard deviation (S_a) of the realized samples, and the ratio between the mean and the standard deviation ($S_a/\bar{a}$) are presented. There exists discrepancy between the sample mean and the reference value. The discrepancy may be

ν	ℓ_{SEK} (eV)	σ_{SEK} (eV^2)	σ_{noise} (eV^2)
1	0.815	1.47×10^{-4}	2.11×10^{-4}
2	0.655	1.07×10^{-4}	1.11×10^{-4}
3	0.668	3.08×10^{-4}	3.00×10^{-4}
4	0.469	2.39×10^{-4}	2.22×10^{-4}
5	0.470	2.38×10^{-4}	2.22×10^{-4}
6	0.578	3.76×10^{-4}	4.80×10^{-4}
7	0.599	2.20×10^{-4}	2.06×10^{-4}
8	0.594	2.18×10^{-4}	2.06×10^{-4}
9	0.813	1.52×10^{-4}	2.73×10^{-4}

TABLE I. The hyperparameters, i.e., ℓ_{SEK}, σ_{SEK}, and σ_{noise}, for the valence bands of HfCoSb, identified within the ML-II approximation. ν is the index of the corresponding phonon branch.

attributed to two factors. One obvious reason that can be considered is the limited sample size, although this is not the most decisive factor in this case. Rather, the central reason for the discrepancy is that the statistical distributions of these thermoelectric properties are not normal, which will be discussed in more detail later. As shown in Figure 1, the Seebeck coefficients exhibit relatively little dependency on the changes in g_ν^2 values. Similarly, the Seebeck coefficients show small standard deviations in Table II, which are only 3-8% of the corresponding mean values. On the other hand, the electrical conductivity and the power factor exhibit much larger standard deviations, amounting to about 20% of the corresponding mean values. This is probably a natural behavior, since the electrical conductivity is directly proportional to the relaxation time, which is directly affected by the uncertainty in g_ν^2. The power factor is again proportional to the electrical conductivity, and hence experiences a similar level of uncertainty. Overall, the result clearly indicates that we can place more confidence in our predicted values of the Seebeck coefficient than in those of the other properties.

Figure 2 shows the histograms of the thermoelectric properties at two different temperatures (300K and 700K). We additionally present the histograms of the resistivity ($1/\sigma$), which is the inverse of the electrical conductivity. As previously mentioned, the statistical distributions of the Seebeck coefficient, the electrical conductivity, and the power factor are

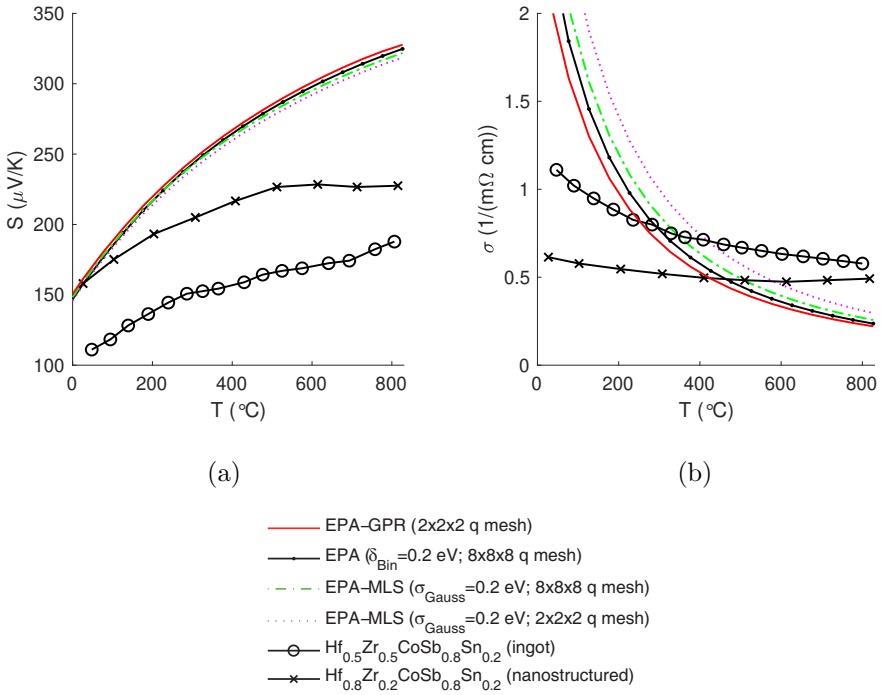


FIG. 1. (Color online.) The Seebeck coefficient S and the electrical conductivity σ for p-type HfCoSb as a function of temperature T calculated with the EPA method and the EPA-MLS method. Consult the legend for the condition represented by each curve. Calculations are performed at the carrier concentration $p = 0.06$ per formula unit. The open circles and the crosses show the experimental data for the ingot sample of p-type $Hf_{0.5}Zr_{0.5}CoSb_{0.8}Sn_{0.2}$[22] and the experimental data for the nanostructured sample of p-type $Hf_{0.8}Zr_{0.2}CoSb_{0.8}Sn_{0.2}$[23,24], respectively.

not normal, as clearly seen in Figures 2 (a), (b), and (d). We have also quantitatively tested the normality of these distributions by applying the Jarque-Bera test[30] to each set of realized samples. The Jarque-Bera test checks the null hypothesis that each data set comes from a normal distribution with an unknown mean and variance. The p-value of the test is the probability of observing a test statistic as extreme as, or more extreme than, the observed sample under the null hypothesis. The p-values of our test for the Seebeck coeffcient, the electrical conductivity, and the power factor at two different temperatures had values much less than 1%, clearly rejecting the null hypothesis for these thermoelectric properties. On the other hand, we visually recognize that the resistivity in Figure 2 (c) exhibits distributions very close to normal. Indeed, the p-value of the Jarque-Bera test for the resistivity is 28.2% at 300K and 15.8% at 700K. The resistivity is the inverse of the electrical conductivity and

a	S (μV/K)		σ (1/(mΩ cm))		PF= $S^2\sigma$ (mW/(m K^2))	
T	300K	700K	300K	700K	300K	700K
a_{ref}	161.7	272.3	2.112	0.4916	5.520	3.646
$\bar{a}$	162.4	272.5	2.115	0.4898	5.537	3.625
S_a	12.95	10.45	0.4289	6.910×10^{-2}	1.092	0.4612
$\text{S}_a/\bar{a}$	7.977×10^{-2}	3.835×10^{-2}	0.2028	0.1411	0.1973	0.1272

TABLE II. Statistical properties of the Gaussian process g_ν^2 and its realized samples at two different temperatures (300K and 700K). The statistical properties of three different thermoelectric properties, i.e., the Seebeck coefficient (S), the electrical conductivity (σ), and the power factor (PF= $S^2\sigma$), are provided. For each thermoelectric property (a), the reference value (a_{ref}) obtained from $\bar{\mathbf{f}}_*$ of Eq. (14), the mean ($\bar{a}$) and standard deviation (S_a) of the realized samples, and the ratio between the mean and the standard deviation ($\text{S}_a/\bar{a}$) are presented. All the values except those of $\text{S}_a/\bar{a}$, which are dimensionless, are reported in the unit corresponding to each quantity.

hence can be considered to be roughly proportional to the scattering rate, i.e., τ^{-1}, which is in turn proportional to the value of g_ν^2. Since g_ν^2 follows a multivariate Gaussian statistic in our numerical model, it is more natural for the resistivity to follow a normal statistic, which is indeed the case in our numerical test.

One of the most important statistical properties that are relevant to the fast-screening procedure of thermoelectric materials is the confidence interval of the estimation. The sample statistics can be utilized to provide such information. In Figure 3, the 5%, 50%, and 95% percentiles for the thermoelectric properties of interest are provided, along with the reference curve directly computed from $\bar{\mathbf{f}}_*$ of Eq. (14). The colored range between the 5% and 95% percentiles indicates a confidence interval of 90%. As discussed already, the confidence intervals of the electrical conductivity and the power factor are relatively large. For example, at 300K, the 95% percentile value, i.e., 7.469 mW/(m K^2) and the 5% percentile value, i.e., 4.113 mW/(m K^2), of the power factor deviate by 35% and 25%, respectively, from the reference value, i.e, 5.520 mW/(m K^2). Clearly, the range is still acceptable for fast-screening applications, but one must remain cautious not to place blind faith on the values from computational estimations.

So far, we have considered the uncertainty in the electronic transport properties of the

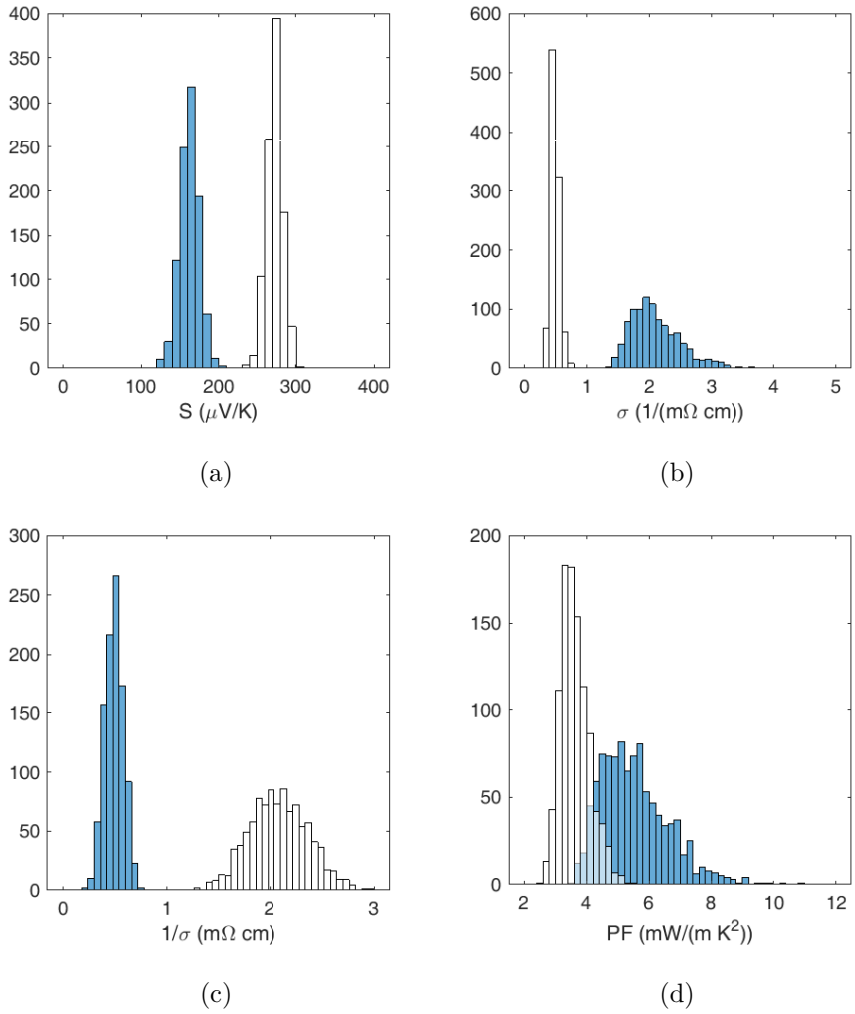


FIG. 2. (Color online.) The histograms of thermoelectric properties at two different temperatures (300K and 700K): (a) the Seebeck coefficient (S); (b) the electrical conductivity (σ); (c) the resistivity ($1/\sigma$); and (d) the power factor (PF). The blue-faced bars and the semi-transparent bars represent data at 300K and at 700K, respectively.

material at a fixed carrier concentration, i.e., $p = 0.06$ per formula unit, but the electronic transport coefficients of materials depend strongly on the carrier concentration[11]. One of the most important objectives of computational prediction is to suggest an optimal carrier concentration for a given composition. The values of g_ν^2 bear certain uncertainty, and hence the predicted optimal carrier concentration will also involve uncertainty. In a previous study[11], it was reported that the values of the optimal carrier concentration maximizing ZT were only about 10% different in average from the values of carrier concentration maximizing PF. Therefore, the PF values from several realized samples at 700K are plotted versus the

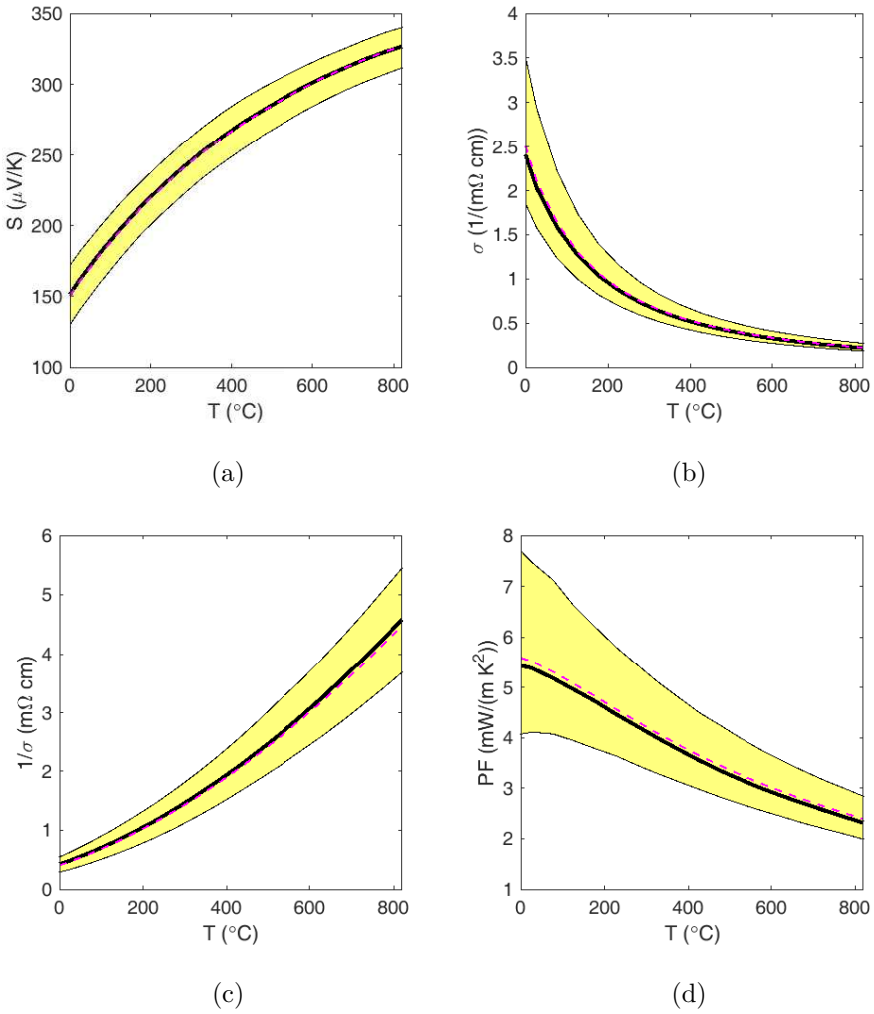
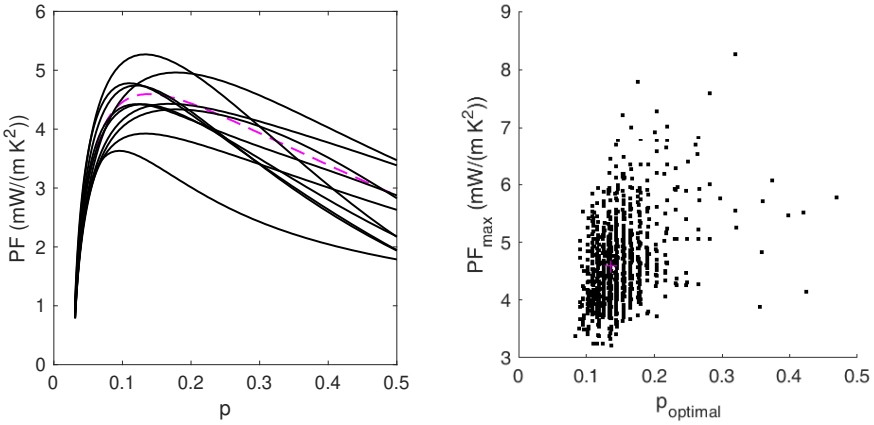


FIG. 3. (Color online.) The 5%, 50%, and 95% percentile curves of thermoelectric properties plotted versus T, presented along with the reference curve directly computed from $\bar{\mathbf{f}}_*$ of Eq. (14): (a) the Seebeck coefficient (S); (b) the electrical conductivity (σ); (c) the resistivity ($1/\sigma$); and (d) the power factor (PF). The thick solid curves represent the 50% percentiles, while the thin solid curves represent the 5% and 95% percentiles. The dashed curves represent the corresponding reference curves. The colored range between the 5% and 95% percentiles represents the confidence interval of 90%.

hole concentration p in Figure 4 (a). The result shows that there exist large variations in the maximum values of PF and the optimal values of p associated, among the chosen samples. Such uncertainty in the values of the optimal carrier concentration and that in the associated PF values should be carefully quantified, and our method can be utilized for serving such a purpose. In Figure 4 (b), a scatter plot showing the maximum value of PF ($\mathrm{PF}_{\max}$) and

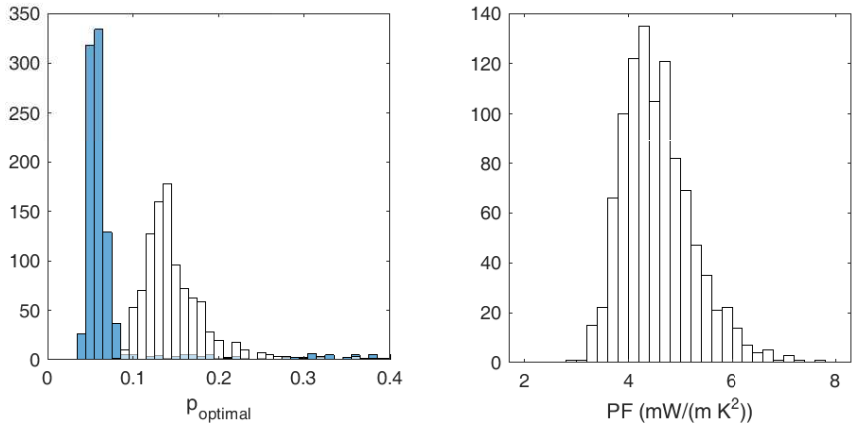


(a) (b)

FIG. 4. (Color online.) Variations in the maximum power factor ($\mathrm{PF}_{\mathrm{max}}$) and the optimal value of hole concentration (p_{optimal}) at 700K: (a) PF versus p for 10 realized samples (solid black) and the reference (dashed magenta) from $\bar{\mathbf{f}}_*$ of Eq. (14); and (b) a scatter plot showing p_{optimal} and the associated maximum PF ($\mathrm{PF}_{\mathrm{max}}$) for each realized sample (black dots) and for the reference (a magenta cross). In both plots, data from too small hole concentrations ($p < 0.03$) have been excluded, since the character of the material changes from a p-type to an n-type there.

associated optimal p value (p_{optimal}) of each realized samples. A fairly large variation is observed, but there is an underlying trend. The samples are scattered around the reference point, where $p_{\mathrm{optimal}} = 0.136$ per formula unit, with an area with high density in the range of $0.1 < p_{\mathrm{optimal}} < 0.2$. The value of $\mathrm{PF}_{\mathrm{max}}$ at this new reference point is 4.595 mW/(m K^2), which is higher than that reported for $p = 0.06$ (Table II).

In Figure 5 (a), we present the histograms of the optimal value of hole concentration (p_{optimal}) at 300K and 700K. At 300K, the most probable value of p_{optimal} turns out to be around 0.06, which was the value employed for our study mentioned above, i.e., Tables I-II and Figures 1-3. On the other hand, the most probable value of p_{optimal} at 700K occurs in between 0.135 and 0.145, which is larger than 0.06. Increase in temperature activates carriers in a wider energy range, and a too low p value may result in a conflict between two different charge carriers, i.e., electrons and holes, resulting in a very low value or even a sign reversal of the Seebeck coefficient at high temperature. Thus, it is natural to find that the most probable p_{optimal} value at 700K is larger than that at 300K. The range of the most probable p_{optimal} at 700K, observed from the histogram, includes p_{optimal} of the



(a) (b)

FIG. 5. (Color online.) The histograms of the optimal value of hole concentration (p_{optimal}) and the PF values associated: (a) p_{optimal} at 300K and 700K; and (b) PF at 700K with $p = 0.14$ per formula unit. The blue-faced bars and the semi-transparent bars represent data at 300K and at 700K, respectively, if there are histograms for both temperatures.

reference point, which is 0.136, observed in Figure 4 (b). A large number of realized samples exhibit maximum values in PF within the range of $0.1 < p_{\text{optimal}} < 0.2$, which confirms our observation made from Figure 4 (b).

In Figure 5 (b), we additionally present the histogram of the PF values with $p = 0.14$ per formula unit, which is the value lying at the center of the range exhibiting the most probable p_{optimal} value ($0.135 < p < 0.145$), at 700K. This plot shows the distribution of possible PF values, which are expected to be achieved if the carrier concentration is optimized at the computationally predicted p value during the synthesis process. The expected PF values are lying mostly in between 3 mW/(m K^2) and 6 mW/(m K^2), exhibiting a significant variation. Such information on the potential uncertainty in the predicted values can be useful for assessing the feasibility of a candidate material.

IV. CONCLUSIONS

We presented the EPA-GPR method where the el-ph coupling matrix is represented as a function of two energies, which is in turn modeled as a Gaussian process. Unlike the other EPA-related methods, the EPA-GPR method is a truly parameter-free estimation method,

since all the hyperparameters in the model can be unambiguously determined within the ML-II approximation. On top of that, the use of a Gaussian process allows us to quantify the uncertainty in the estimated thermoelectric properties.

To demonstrate the effectiveness of the EPA-GPR method, we applied it to a p-type half-Heusler compound, i.e., HfCoSb. Our numerical results clearly exhibit the advantages of the method. In particular, we note that the estimated power factor can vary up to about 35% at room temperature within a confidence level of 90%, which is acceptable for fast-screening applications but still requires a certain level of caution in fast-screening applications. Overall, the information on the potential uncertainty in computational prediction can be valuable in future decision-making processes of the research and development of new TE materials.

ACKNOWLEDGMENTS

The first, second, and third authors were mainly supported by Solvay SA, an advanced materials and specialty chemicals company, through an Ewha-Solvay collaboration agreement, during the course of the study reported in this article. The forth and fifth authors were partly supported by the U.S. Department of Energy under Award DE-EE0004840.

---

* dhwee@ewha.ac.kr; corresponding author.

[1] S. Twaha, J. Zhu, Y. Yan, and B. Li, Renewable and Sustainable Energy Reviews **65**, 698 (2016).

[2] D. A. Broido, M. Malorny, G. Birner, N. Mingo, and D. A. Stewart, Applied Physics Letters **91**, 231922 (2007).

[3] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo, Phys. Rev. X **4**, 011019 (2014).

[4] A. van Roekeghem, J. Carrete, C. Oses, S. Curtarolo, and N. Mingo, Phys. Rev. X **6**, 041061 (2016).

[5] J. M. Ziman, *Principles of the Theory of Solids*, 2nd ed. (Cambridge University Press, 1972).

[6] G. K. Madsen and D. J. Singh, Computer Physics Communications **175**, 67 (2006).

[7] M. Bernardi, D. Vigil-Fowler, J. Lischner, J. B. Neaton, and S. G. Louie, Phys. Rev. Lett. **112**, 257402 (2014).

[8] J. Yan, P. Gorai, B. Ortiz, S. Miller, S. A. Barnett, T. Mason, V. Stevanovic, and E. S. Toberer, Energy Environ. Sci. **8**, 983 (2015).

[9] A. J. Hong, L. Li, R. He, J. J. Gong, Z. B. Yan, K. F. Wang, J.-M. Liu, and Z. F. Ren, Scientific Reports **6**, 22778 (2016).

[10] J. Noffsinger, F. Giustino, B. D. Malone, C.-H. Park, S. G. Louie, and M. L. Cohen, Computer Physics Communications **181**, 2140 (2010).

[11] G. Samsonidze and B. Kozinsky, Advanced Energy Materials **8**, 1800246 (2018).

[12] S. Bang, J. Kim, D. Wee, G. Samsonidze, and B. Kozinsky, Materials Today Physics **6**, 22 (2018).

[13] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, 2006).

[14] F. Giustino, M. L. Cohen, and S. G. Louie, Phys. Rev. B **76**, 165108 (2007).

[15] W. Li, Phys. Rev. B **92**, 075405 (2015).

[16] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, Rev. Mod. Phys. **73**, 515 (2001).

[17] D. Levin, Math. Comput. **67**, 1517 (1998).

[18] C. E. Rasmussen and H. Nickisch, J. Mach. Learn. Res. **11**, 3011 (2010).

[19] A. G. Wilson and H. Nickisch, in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15 (JMLR.org, 2015) pp. 1775–1784.

[20] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, Journal of Physics: Condensed Matter **21**, 395502 (19pp) (2009).

[21] MATLAB, *Statistics and Machine Learning Toolbox^{TM} User's Guide* (The MathWorks, Inc., Natick, Massachusetts, 2018).

[22] S. R. Culp, J. W. Simonson, S. J. Poon, V. Ponnambalam, J. Edwards, and T. M. Tritt, Applied Physics Letters **93**, 022105 (2008).

[23] X. Yan, W. Liu, H. Wang, S. Chen, J. Shiomi, K. Esfarjani, H. Wang, D. Wang, G. Chen, and Z. Ren, Energy & Environmental Science **5**, 7543 (2012).

[24] X. Yan, G. Joshi, W. Liu, Y. Lan, H. Wang, S. Lee, J. W. Simonson, S. J. Poon, T. M. Tritt,

G. Chen, and Z. F. Ren, Nano Letters **11**, 556 (2011).

[25] A. Page, P. Poudeu, and C. Uher, Journal of Materiomics **2**, 104 (2016).

[26] F. Casper, T. Graf, S. Chadov, B. Balke, and C. Felser, Semiconductor Science and Technology **27**, 063001 (2012).

[27] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[28] D. Vanderbilt, Phys. Rev. B **41**, 7892 (1990).

[29] A. D. Corso, Computational Materials Science **95**, 337 (2014).

[30] C. M. Jarque and A. K. Bera, Economics Letters **6**, 255 (1980).